Response to NIST AI RMF Initial Draft

28 April 2022

Elham Tabassi, Chief of Staff, Information Technology Laboratory
National Institute of Standards and Technology (NIST)
100 Bureau Drive, Gaithersburg, MD 20899

Subject: NIST AI Risk Management Framework Initial Draft

Via email to AIframework@nist.gov

To Ms. Tabassi, and the entire NIST team developing the AI Risk Management Framework,

Thank you for the invitation to submit comments in response to the Initial Draft of the NIST AI Risk Management Framework (AI RMF or Framework). We offer the following submission for your consideration.

We are researchers affiliated with UC Berkeley, with expertise on AI research and development, safety, security, policy, and ethics. We previously submitted responses to NIST in September 2021 on the NIST AI RMF Request For Information (RFI), and in January 2022 on the AI RMF Concept Paper.

In the following sections, we provide in-depth comments, first regarding the questions posed by NIST in the AI RMF Initial Draft, and then on specific passages in the NIST AI RMF Initial Draft.

Thank you again for the opportunity to comment on the AI RMF Initial Draft. If you need additional information or would like to discuss further, please contact Anthony Barrett at anthony.barrett@berkeley.edu. In any case, we look forward to further engagement with NIST as you proceed on the AI RMF development process.

Our best,

Anthony Barrett, Ph.D., PMP
Visiting Scholar
AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley

Dan Hendrycks
Ph.D. Candidate
Berkeley AI Research Lab, UC Berkeley

Jessica Newman
Director

AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley
Co-Director
AI Policy Hub, UC Berkeley

Mark Nitzberg, Ph.D.
Executive Director
Center for Human-Compatible AI, UC Berkeley
Head of Strategic Outreach
Berkeley AI Research Lab, UC Berkeley

Brandie Nonnecke, Ph.D.
Director
CITRIS Policy Lab, CITRIS and the Banatao Institute, UC Berkeley
Co-Director
AI Policy Hub, UC Berkeley

Richmond Y. Wong, Ph.D.
Postdoctoral Scholar
Center for Long-Term Cybersecurity, UC Berkeley

Our comments on questions posed by NIST in the AI RMF Initial Draft

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

Response/Comment: Overall, we do not believe that the AI RMF sufficiently covers and addresses AI risks, especially systemic and societal risks, nor does it have sufficient specificity for various use cases. However, we believe it has the potential to do so if expanded out, and that the question about the right level of specificity will depend in part on the draft Practice Guide, which has not been released at this time.

2. Whether the AI RMF is flexible enough to serve as a continuing resource considering evolving technology and standards landscape.

Response/Comment: Our current sense is that the AI RMF would be flexible enough to serve as a continuing resource, especially with frequent updates to the Practice Guide and Profiles.

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

Response/Comment: Overall, we do believe that the AI RMF has potential to enable decisions about how an organization can increase understanding of, communication about, and efforts to manage AI risks. However, again, much will depend on the draft Practice Guide, which has not been released at this time. We would expect the Practice Guide to include directly, or include via reference, a wide array of example pitfalls, concerns, and remediations. Additional recommendations on specific methodologies within the main AI RMF document would also be very helpful.

4. Whether the functions, categories, and subcategories are complete, appropriate, and clearly stated.

Response/Comment: The functions, categories, and subcategories need greater detail to be practical for a wide variety of potential users. We provide several more specific comments addressing this question in the following section “Our comments on specific passages in the NIST AI RMF Initial Draft”, under “Page 16, Table 1, Category ID 2”; “Page 16, Table 1, Category ID 3”; “Page 16, Table 1, Category ID 4”; “Page 17, Table 2, Category ID 2, Second Subcategory”; etc.

5. Whether the AI RMF is in alignment with or leverages other frameworks and standards such as those developed or being developed by IEEE or ISO/IEC SC42.

Response/Comment: Our current sense is that the AI RMF is broadly in alignment with other frameworks and standards such as those developed or being developed by ISO/IEC JTC 1 SC42. However, further details about the AI RMF are needed, and many of the other frameworks and standards are also being continually developed, so this issue should be revisited.

6. Whether the AI RMF is in alignment with existing practices, and broader risk management practices.

Response/Comment: Overall, we do believe that the AI RMF is in alignment with many current best practices, including risk management practices. We provide more specific comments in the following.

7. What might be missing from the AI RMF.

Response/Comment 7A:
We believe that something missing from the Initial Draft of the AI RMF is clearer discussion of potential for systemic or even catastrophic impacts to individuals and society. We agree with the

statements on p. 6 of the Initial Draft that examples of potential harms from AI systems include systemic risks such as “large scale harms to the financial system or global supply chain”, and long term risks as follows: “Some AI risks … may be latent at present but may increase in the long term as AI systems evolve.” However, the Initial Draft does not seem to have a statement clearly corresponding to the following passage from the AI RMF Concept Paper: “...Managing AI risks presents unique challenges. An example is the evaluation of effects from AI systems that are characterized as being long-term, low probability, systemic, and high impact. Tackling scenarios that can represent costly outcomes or catastrophic risks to society should consider: an emphasis on managing the aggregate risks from low probability, high consequence effects of AI systems, and the need to ensure the alignment of ever more powerful advanced AI systems.” (NIST 2021a, p.1) In the wake of significant advances of AI systems such as BERT, CLIP, GPT-3, DALL-E 2, and PaLM, it is vitally important for the AI RMF to prepare teams for addressing the possibility of both transformative benefits and catastrophic risks of these increasingly multi-purpose or general-purpose AI that can serve as AI platforms underpinning many end-use applications. Such advanced AI models often have qualitatively distinct properties compared to narrower models, such as the potential to be applied to many sectors at once, and emergent properties that can provide unexpected capabilities but also unexpected risks of adverse events. These models could present corresponding catastrophic risks to society, e.g. of correlated robustness failures across multiple high-stakes application domains (Bommasani et al. 2021 pp. 115-116).

We recommend that NIST more clearly adapt or insert statements from that AI RMF Concept Paper passage into the AI RMF Section 4 (Framing Risk), perhaps specifically in Section 4.2 (Challenges for AI Risk Management). We believe that it would be in the interests of all stakeholders, including AI developers, for the AI RMF to clearly aim to constructively prompt early, proactive consideration of these risk management issues. As Jihao Chen of Parity AI and Richard Mallah of the Future of Life Institute both noted in the NIST AI RMF Workshop 2 (NIST 2022), identifying and addressing a risk earlier instead of later helps to maximize benefits and minimize costs of managing that risk. Moreover, there is precedent for NIST framework guidance prompting risk assessment considering potentially catastrophic impacts: the NIST Cybersecurity Framework guidance on risk assessment points to NIST SP 800-53 RA-3, which in turn references NIST SP 800-30; the impact assessment scale in Table H-3 of SP 800-30 includes criteria for rating an expected impact as a “catastrophic adverse effect” to individuals, organizations, or a society (NIST 2012, NIST 2018, NIST 2020).

Suggested Change 7A:
In Section 4.2 or elsewhere, we recommend adding statements that more clearly correspond to the following passage (or perhaps just insert the entire passage) from p.1 of the AI RMF Concept Paper: “...Managing AI risks presents unique challenges. An example is the evaluation of effects from AI systems that are characterized as being long-term, low probability, systemic, and high impact. Tackling scenarios that can represent costly outcomes or catastrophic risks to society should consider: an emphasis on managing the aggregate risks from low probability, high consequence effects of AI systems, and the need to ensure the alignment of ever more powerful advanced AI systems.”

Response/Comment 7B:
In addition, the current framing of AI risks and characteristics of trustworthy AI described in Figure 3 and elsewhere may be missing important details and nuance. It seems unclear why in Figure 3 and elsewhere, NIST only selected the three guiding principles of fairness, accountability, and transparency and not others. For example, the NIST Taxonomy of AI Risk (NIST 2021b, p. 8) notes that the OECD principles include "traceability to human values" and EU principles include "human agency and oversight" and "environmental and societal well-being."

More broadly, the list of known risks and characteristics of trustworthy AI as shown in Figure 3 is not comprehensive. There are many additional characteristics that may inform the realization of trustworthy AI. Although these cannot all reasonably be added here, it should be acknowledged that more detail and nuance is available elsewhere. (See, e.g., discussion in CLTC forthcoming.) It would be valuable for NIST to provide guidance on how organizations can incorporate additional guiding principles and/or characteristics as part of their use of the AI RMF. Lastly, the split between technical and socio-technical risks is problematic. We provide further discussion and recommendations about this point in the following comment under "Page 7, Lines 35-37; Page 8, Figure 3 and elsewhere in Section 5."

Suggested Change 7B:
In Section 5 or elsewhere, we recommend providing an explanation of why NIST only selected the three guiding principles of fairness, accountability, and transparency and not others. We also recommend providing guidance on how organizations can incorporate additional guiding principles and/or characteristics as part of their use of the AI RMF.

References:
Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, Bernstein MS, Bohg J, Bosselut A, Brunskill E, Brynjolfsson E, Buch S, Card D, Castellon R, Chatterji N, Chen A, Creel K, Davis JQ, Demszky D, Donahue C, Doumbouya M, Durmus E, Ermon S, Etchemendy J, Ethayarajh K, Fei-Fei L, Finn C, Gale T, Gillespie L, Goel K, Goodman N, Grossman S, Guha N, Hashimoto T, Henderson P, Hewitt J, Ho DE, Hong J, Hsu K, Huang J, Icard T, Jain S, Jurafsky D, Kalluri P, Karamcheti S, Keeling G, Khani F, Khattab O, Kohd PW, Krass M, Krishna R, Kuditipudi R, Kumar A, Ladhak F, Lee M, Lee T, Leskovec J, Levent I, Li XL, Li X, Ma T, Malik A, Manning CD, Mirchandani S, Mitchell E, Munyikwa Z, Nair S, Narayan A, Narayanan D, Newman B, Nie A, Niebles JC, Nilforoshan H, Nyarko J, Ogut G, Orr L, Papadimitriou I, Park JS, Piech C, Portelance E, Potts C, Raghunathan A, Reich R, Ren H, Rong F, Roohani Y, Ruiz C, Ryan J, Ré C, Sadigh D, Sagawa S, Santhanam K, Shih A, Srinivasan K, Tamkin A, Taori R, Thomas AW, Tramèr F, Wang RE, Wang W, Wu B, Wu J, Wu Y, Xie SM, Yasunaga M, You J, Zaharia M, Zhang M, Zhang T, Zhang X, Zhang Y, Zheng L, Zhou K, and Liang P (2021), On the Opportunities and Risks of Foundation Models. *arXiv*, https://arxiv.org/abs/2108.07258

CLTC (forthcoming). Identifying Properties of Trustworthiness for a Spectrum of AI Models & Applications. UC Berkeley Center for Long Term Cybersecurity.

NIST (2012) Guide for Conducting Risk Assessments, SP 800-30 Rev. 1. National Institute of Standards and Technology, https://csrc.nist.gov/publications/detail/sp/800-30/rev-1/final

NIST (2018) Framework for Improving Critical Infrastructure Cybersecurity, Version 1.1. National Institute of Standards and Technology, https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.04162018.pdf

NIST (2020) Security and Privacy Controls for Information Systems and Organizations, SP 800-53 Rev. 5. National Institute of Standards and Technology, https://csrc.nist.gov/publications/detail/sp/800-53/rev-5/final

NIST (2021a) AI Risk Management Framework Concept Paper. National Institute of Standards and Technology. https://www.nist.gov/document/airmfconceptpaper.

NIST (2021b) Draft - Taxonomy of AI Risk. National Institute of Standards and Technology. https://www.nist.gov/system/files/documents/2021/10/15/taxonomy_AI_risks.pdf

NIST (2022) Building the NIST AI Risk Management Framework: Workshop #2. National Institute of Standards and Technology. https://www.nist.gov/news-events/events/2022/03/building-nist-ai-risk-management-framework-workshop-2

8. 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.

Response/Comment: The draft Practice Guide has not been released at this time. It would be useful if it included, among other things, guidance on how to identify possible harms and risks, as well as reference lists of examples of unintended harms that AI systems can cause or have caused.

9. Others?

Note: This first draft does not include Implementation Tiers as considered in the concept paper. Implementation Tiers may be added later if stakeholders consider them to be a helpful feature in the AI RMF. Comments are welcome.

Response/Comment: Our current sense is that for many AI RMF users, Implementation Tiers will not have great value. Moreover, omitting Implementation Tiers may help avoid confusion about differences between NIST Implementation Tiers and EU AI Act risk tiers.

Our comments on specific passages in the NIST AI RMF Initial Draft

Page 1, Lines 16-17

Response/Comment: The three principles mentioned here (i.e., accountability, fairness, and equity) do not match those listed in the Table of Contents (and in Figures 3 and 4): fairness, accountability, and transparency.

Suggested Change: We recommend ensuring consistency of the terms used throughout the text and providing explanations as to why specific terms were selected.

Page 4, Figure 1, and Lines 8-14 and throughout Section 3

Response/Comment: The term “AI System Stakeholders” seems confusing and potentially worth splitting into more specific groups. As NIST notes in its Section 3 discussions, for many AI systems there will be a variety of key stakeholder groups beyond the much smaller set defined as “AI System Stakeholders.” Furthermore, all parties represented in the diagram are stakeholders of the AI system even if they are not responsible for its development. In many cases, it will also be important to distinguish between AI system developers and deployers. Moreover, AI system development often comprises a value chain across multiple entities with corresponding responsibilities such as communication about AI system limitations and risks, so it could be useful if “developers” could be broken further, perhaps into "upstream developers" (e.g. OpenAI for GPT-3) and "downstream developers" (e.g. for applications building on GPT-3).

Conversely, in some cases, a particular stakeholder could be placed in more than one group, e.g. if a developer/deployer is also an operator. It could also be better to frame the stakeholder "rings" on the "role" or "activity" being conducted by a stakeholder. This would allow for an individual stakeholder to be placed in more than one ring of the diagram.

Suggested Change: Consider refinements to the terms and depictions of the AI RMF key stakeholder groups, e.g. to more easily distinguish between developers and deployers (or even between upstream and downstream developers in a value chain), or to allow a particular stakeholder to be placed in more than one group.

Page 4, Lines 8-14

Response/Comment: A role not mentioned in the AI system stakeholder group is “data workers”, sometimes known as “data annotators”, “data janitors”, or other names (c.f. Irani 2015, Miceli et al 2020, Pine and Bossen 2020). Often distinct from design and development teams, these workers are often outsourced or work alongside domain experts in locations where AI systems

are deployed, yet their decisions can have ramifications for the learning behaviors of AI systems.

Suggested Change: Consider including “data workers”, “data annotators,” or a similar term in the list of AI system stakeholders.

References:
Irani L (2015) Justice for “Data Janitors”. https://www.publicbooks.org/justice-for-data-janitors/

Miceli M, Schuessler M, and Yang T (2020) Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 115 (October 2020). https://doi.org/10.1145/3415186

Pine KH, and Bossen C (2020) Good organizational reasons for better medical records: The data work of clinical documentation integrity specialists. *Big Data & Society*. https://doi.org/10.1177/2053951720965616

Page 5, Lines 28-29

Response/Comment: Adaptive strategies are often appropriate as part of managing risks with large and irreducible uncertainties. However, adaptive strategies alone are inadequate for managing risks of events that have already been identified as presenting hazards of catastrophic and irreversible effects if those events occur. Moreover, it does not seem accurate to imply that the AI RMF does not enumerate known risks in advance, because Section 6.1 and 6.2 of the Initial Draft clearly provides procedures for identifying known risks.

Suggested Change: Revise the sentence in lines 28-29 from "Additionally, this framework is designed to be responsive to new risks as they emerge rather than enumerating all known risks in advance" to "Additionally, this framework is designed to be responsive to new risks as they emerge rather than assuming a complete understanding of all potential risks in advance."

Page 6, Figure 3

Response/Comment: It is unclear whether environmental impacts such as emissions from AI model training energy consumption would be under impacts to “systems” or accounted for in some other way. It is also unclear whether societal impacts, such as those that involve interactions between multiple systems, are currently represented.

Suggested Change: We recommend adding a fourth column to Figure 3, titled “Harms to society or the environment”. Or, at the very least, we recommend clarifying whether environmental impacts and other societal-level impacts would be under impacts to “systems” or accounted for in some other way.

Page 6, Lines 12-13

Response/Comment: The intended meaning seems unclear for this sentence: "Some AI risks may have a low probability in the short term but have a high likelihood for adverse impacts."

Suggested Change: We recommend changing the sentence to "Some AI risks may have a low probability in the short term but could cause very high adverse impacts if they occur." That would seem consistent with the valuable discussion of high-consequence, low-probability risks in lines 13-17 of the NIST AI RMF Concept Paper (NIST 2021a).

References:
NIST (2021a) AI Risk Management Framework Concept Paper. National Institute of Standards and Technology. https://www.nist.gov/document/airmfconceptpaper.

Page 7, Lines 3-6

Response/Comment: We believe it is valuable that this passage of the Initial Draft suggests that management decision-making include consideration of whether AI systems present unacceptable risks, and "whether an AI system should be designed, developed or deployed at all."

Suggested Change: We recommend retaining and expanding upon these passages in the Initial Draft, e.g. to add a statement that more clearly encourages organizations to consider entirely avoiding AI systems that pose unacceptable risks to rights, values, or safety.

Page 7, Lines 7-21

Response/Comment: This paragraph discusses the idea that risk thresholds could be set by AI system owners, organizations, industries, communities, and/or regulators. It would be helpful to have example risk thresholds or values to facilitate alignment among stakeholders. In doing so, a shared understanding of types of risks and appropriate risk mitigation strategies may be helpful.

Suggested Change: In the next AI RMF draft, and/or in the Practice Guide, we recommend providing example risk thresholds, e.g. perhaps drawing on implicit thresholds for unacceptable risks in the draft EU AI Act.

Page 7, Lines 32-33

Response/Comment: It could be helpful to expand on this statement (“Small to medium-sized organizations face different challenges in implementing the AI RMF than large organizations”). One reason would be to explain NIST's understanding of potential differences in the needs of various industry actors, as NIST invites feedback on how best to support industry in applying the AI RMF.

Suggested Change: We recommend expanding on this statement, to explain NIST's understanding of potential differences between the needs of small to medium-sized organizations and large organizations.

Page 7, Lines 35-37; Page 8, Figure 3 and elsewhere in Section 5

Response/Comment: The taxonomy includes: technical characteristics, socio-technical characteristics, and guiding principles. This comes from NIST's taxonomy of AI risks (NIST 2021b). However, the differences between “technical” and “socio-technical” characteristics are not at all clear-cut. In fact, all three categories are socio-technical; for each of these categories some degree of “human judgment must be employed when deciding on the specific metrics and the precise threshold values for these metrics” as mentioned in Section 5.2 regarding socio-technical characteristics. At the very least, the considerable overlaps should be noted. Potentially, the categories are simply not helpful.

Suggested Change: We recommend, at minimum, when describing the characteristics and the taxonomy, adding a note such as the following: “To some extent, there are important socio-technical aspects of each of the characteristics we have classified as technical characteristics, socio-technical characteristics, and guiding principles. For each of these, some degree of human judgment must be employed when deciding on the specific metrics and the precise threshold values for these metrics.”

References:
NIST (2021b) Draft - Taxonomy of AI Risk. National Institute of Standards and Technology. https://www.nist.gov/system/files/documents/2021/10/15/taxonomy_AI_risks.pdf

Page 10, Lines 10-11

Response/Comment: The statement “Robustness contributes to sensitivity analysis in the AI risk management process” seems backward, or at least confusing.

Suggested Change: We recommend changing this sentence to either "Sensitivity analysis contributes to robustness" or "One broad category of techniques for evaluating robustness is sensitivity analysis."

Page 10, Line 12

Response/Comment: It seems not entirely clear how “Resilience or ML Security” differs from “Robustness”.

Suggested Change: At minimum, consider changing the term “Resilience or ML Security” to “Security and Resilience,” to emphasize the role of security concerns for this characteristic. Also, consider adding more explanation and guidance on how to address potential overlaps or gaps between robustness, resilience and security concerns.

Page 10, Lines 15-16

Response/Comment: The phrase here, “unexpected or adversarial use of the model or data” seems to point to consideration of potential abuse and/or misuse cases. Mentioning those terms here could helpfully prompt consideration of abuse and/or misuse cases. That could also build on best practices for consideration of adversarial misuse potential of an AI system such as in Microsoft (2021), and/or related software development guidance such as OWASP (2021).

Suggested Change: Consider changing "use" in this passage to "use (or abuse/misuse)".

References:
Microsoft (2021) Foundations of assessing harm. Microsoft, https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/harms-modeling/

OWASP (2021) Abuse Case Cheat Sheet. OWASP, https://cheatsheetseries.owasp.org/cheatsheets/Abuse_Case_Cheat_Sheet.html

Page 13, Line 26; and Page 15, Table 1, Category ID 1

Response/Comment: These passages include the following phrases when discussing an AI system: "intended use case", “intended purpose”, and “use”. We believe there can be drawbacks in employing singular terms such as "use", "use case", “purpose”, “task”, etc. in ways that could suggest only considering a single intended use of an AI system. AI systems can have multiple uses and it is worth identifying these as part of the Map function, e.g. to enable policies disallowing specific uses that would present unacceptable risks. The drawbacks of assuming a single intended “use” would be especially important for increasingly general-purpose AI that can be employed in many end-use applications. For any AI system, and especially for increasingly multi-purpose or general-purpose AI systems such as BERT, CLIP, and GPT-3, focusing on a single intended “use” could overlook many important beneficial opportunities as well as risks of adverse events.

Suggested Change: We recommend that in these passages and throughout AI RMF documents, NIST generally either employ terminology such as AI system “uses” or “use cases” instead of “use” (and “purposes” instead of “purpose”, “tasks” instead of “task”, etc.) or include related notes, to avoid implying that all AI systems would have a single intended use.

Page 16, Table 1, Category ID 2

Response/Comment: The subcategories under the category “Classification of AI system is performed” do not currently appear to include documentation of additional important aspects of the design approach, and expected outputs and behavior.

Suggested Change: Consider adding two subcategories: First, “Information about the design of the AI system, including the choices of model architecture and training procedures (or equivalents), and reasons for those choices, are documented.” Second, “Expected outputs and behavior, including performance metrics and limitations of the AI system are documented.”

Page 16, Table 1, Category ID 3

Response/Comment: The first two subcategories under the category “AI capabilities, targeted usage, goals, and expected benefits and costs over status quo are understood” should be edited because the intended system behavior may include both benefits and costs.

Suggested Change: We recommend changing the first subcategory to say: “The intended system behavior and uses are examined, including associated benefits and costs.” A second should be added to say “Potential system behavior and uses are examined, including associated benefits and costs.” The third should say “Errors or unintended system behavior or uses are examined, including associated benefits and costs.” An additional subcategory should also be added that says “Expected or potential shifts to the application scope or operational context over time are documented.”

Page 16, Table 1, Category ID 4

Response/Comment: This category (“Risks and harms to individual, organizational, and societal perspectives are identified”) only discusses risk identification, not additional analysis of identified risks. However, the subcategories discuss additional steps of risk analysis, including: consequence analysis, i.e. elucidating the magnitude of potential harms via qualitative or quantitative analysis; estimating probabilities, i.e. assessing the likelihood of potential harms; and risk-benefit analysis, i.e. comparing risks and benefits, and confirming (in a reasonably objective fashion) that "benefits of the AI system outweigh the risks".

Suggested Change: We recommend either revising the category passage from "Risks and harms ... are identified" to "Risk and harms ... are identified and analyzed", or moving all of the subcategories on risk analysis steps beyond risk identification (i.e. on consequence analysis, estimating probabilities, and risk-benefit analysis) under the "Systems are evaluated" category of the Measure function.

Page 16, Table 1, Category ID 4, First Subcategory

Response/Comment: Even at the subcategory level, the description here of risk identification activities and outcomes still seems relatively high level in comparison to the specific items that should be part of risk identification. For example, it's not obvious from this subcategory passage, nor from the material on the technical characteristic "Resilience or ML Security", that identification of potential misuse/abuse cases should be a routine part of risk identification. This should be clarified in the next AI RMF and/or in the Practice Guide. In addition, the term "understood" should be clarified to include "identified".

In addition, the term "potential users" seems to overlook potential impacts to other individuals and groups besides users. The word "or" in the passage "potential users, the organizations, or society as a whole" also may be interpreted as suggesting the AI RMF procedures can be applied to identify impacts to potential users, or organizations, or society, but not two or more of those.

Suggested Change: We recommend changing "understood" to either "identified" or "identified and understood". Also consider adding sufficient detail to a subcategory here, and/or to the material on the technical characteristic "Resilience or ML Security", so that it becomes clear that the identification of potential misuse/abuse cases should be a routine part of risk identification.

In addition, we recommend changing "potential users, the organizations, or society as a whole" to "potential users and other individuals, groups, organizations and society as a whole (as appropriate)".

Page 17, Table 2, Category ID 2, Second Subcategory

Response/Comment: This subcategory (“Mechanisms for tracking identified risks over time are in place, particularly if potential risks are difficult to assess using currently available measurement techniques, or are not yet available”) seems potentially valuable as part of a proactive effort to manage risks that are not yet easily assessed. It also seems likely actionable for many key AI development organizations employing enterprise risk management, e.g. by tracking identified risks using a risk register. (For more on risk registers, see e.g., ISO Guide 73 Section 3.8.2.4 and PMI 2017 p. 417.)

Suggested Change: We recommend expanding upon this draft text in the next version and/or in practice guides, e.g. to suggest tracking identified risks using a risk register, and to perform periodic reviews and updates based on newly available information.

References:
ISO (2009) ISO Guide 73:2009, Risk management — Vocabulary. https://www.iso.org/obp/ui/#iso:std:iso:guide:73:ed-1:v1:en

PMI (2017) Guide to the Project Management Body of Knowledge, Sixth Edition. Project Management Institute, Newtown Square, PA

Page 18, Lines 8-10

Response/Comment: We agree with this passage, that governance should address supply chains, including third-party software or hardware systems and data as well as internally developed AI systems. (The original passage was missing the word “as”.)

Suggested Change: We recommend considering edits to other sections to clarify that supply chains and third party software should be considered in each function, e.g. to identify supply chain risks as part of Map function activities. We also recommend adapting or referring readers to NIST resources on supply chain risk management, such as NIST SP 800-161 (NIST 2015).

In addition, we recommend changing “as well” to “as well as” in this passage.

References:
NIST (2015) Supply Chain Risk Management Practices for Federal Information Systems and Organizations. National Institute of Standards and Technology, https://csrc.nist.gov/publications/detail/sp/800-161/final

Page 19, Table 4, Category ID 1, Third Subcategory

Response/Comment: This subcategory (“Methods for ensuring all dimensions of trustworthy AI are embedded into policies, processes, and procedures”) seems extremely vague and broad, and would benefit from greater specificity.

Suggested Change: We recommend greater specificity, for example by adding: 1: “Ongoing monitoring and periodic review of the AI system and its outcomes are planned, with responsibilities clearly defined.” (i.e. not just monitoring the risk management process, but also the AI system.) 2. “Documentation practices for the AI system(s) are established.” 3. “AI incident reporting processes are established.” 4. “User communication and redress mechanisms are established.”

Page 19, Table 4, Category ID 4

Response/Comment: We believe this category (“Teams are committed to a culture that considers and communicates risk”) and its current draft subcategories seem valuable. However, the subcategories do not clearly include communicating with users and other stakeholders.

Suggested Change: Consider adding a subcategory that says: “Risk communication practices are identified not only for internal communication, but also for communication with users and other stakeholders.” Or consider adding that language to the current draft subcategory "Teams are encouraged to consider and document the impacts of the technology they design and to develop and communicate about these impacts more broadly."