April 27, 2022

Response of The MITRE Corporation to the NIST Request for Information on the AI Risk Management Framework: Initial Draft (released March 17, 2022)

Approved for Public Release; Distribution Unlimited. Public Release Case Number 22-1338

For additional information about this response, please contact:
Mike Hadjimichael or Michael Garris
The MITRE Corporation
7515 Colshire Drive
McLean, VA 22102-7539
{mikeh,mgarris}@mitre.org
703.983.6000

<<This page is intentionally blank.>>

Contents

Introduction

The MITRE Corporation is pleased to respond to the Request for Information on the "NIST AI Risk Management Framework: Initial Draft" released by the National Institute of Standards and Technology (NIST) on March 17, 2022.

As a not-for-profit organization, the MITRE Corporation works in the public interest to tackle difficult problems that challenge the safety, stability, security, and well-being of our nation through the operation of multiple federally funded research and development centers and labs and through participation in public-private partnerships. Working across federal, state, and local governments — as well as industry and academia — gives MITRE a unique vantage point. MITRE works to discover new possibilities, create unexpected opportunities, and lead by pioneering research for the public good to bring innovative ideas into existence in areas such as artificial intelligence (AI), intuitive data science, quantum information science, health informatics, policy and economic expertise, trustworthy autonomy, cyber threat sharing, and cyber resilience.

MITRE has a long history of partnering with federal agencies to apply the best elements of AI and machine learning (ML) while developing and supporting ethical guardrails to protect people and their personal data. Our team is committed to anticipating and solving future needs that are vital to the success and safety of the public and the country.

In the following pages, we offer thoughts on actions NIST should take to strengthen the draft AI Risk Management Framework (RMF) document. MITRE chose to focus on areas drawing from its decades of experience in supporting the government in its adoption of AI technologies. In our response, we highlight four major themes, but also address a large variety of lesser issues. In particular, we've identified these themes:

1. Look beyond statistical machine learning.
2. Look beyond the system component level.
3. Align and disambiguate with the existing NIST RMF.
4. Include greater emphasis on risks arising from Adversarial ML.

MITRE values the opportunity to contribute to this important work, and we are eager to engage further with NIST and the community it is leading.

Requested Feedback

This document primarily responds to NIST's request for:

- What might be missing from the AI RMF?
- Whether the functions, categories, and subcategories in the draft AI RMF are complete, appropriate, and clearly stated.

I. Major Themes

Look beyond statistical machine learning (ML)

Recommendation: NIST should not overly-constrain or limit the scope of this AI RMF to data-driven ML-based technologies and systems.

The callout box on page 2, line 7 states that this document applies to "algorithmic processes that learn from data in an automated or semi-automated manner." While data-driven, statistical ML represents much of today's AI deployments, it is but one type of AI belonging to a much larger constellation of methodologies and techniques, and the next wave of AI is anticipated to evolve beyond heavy dependence on large amounts of labeled and unlabeled data.[1] To this end, NIST should do its best to lead the development of a framework that is able to manage the risks of many types of AI and is reasonably future proof so as to not quickly lose relevance or become obsolete. It is suggested that the scope of this AI RMF and its supporting language be framed around AI for decision assistance, decision making, and action taking/control and the scoping text in the callout box on page 2 be correspondingly changed or removed.

Accepting and acting upon this recommendation will have ripple effects across the document where at times ML is currently used synonymously with AI - rather than using ML as an exemplar type of AI and calling out particular risks that at times arise with data-driven, statistical ML. The use of "ML Security" as a synonym for AI resilience is a clear example that should be changed. Suggested edits are provided later in this response to make these types of changes.

Recommendation: The 5.1 Technical Characteristics section should be expanded to capture misspecification risks from AI, in addition to risks from deficiencies in accuracy, reliability, robustness, and resilience.

The 5.1 Technical Characteristics section in the current draft is effective for evaluating risk from most current ML implementations. However, it is too narrow to fully describe risks from AI systems that are "accurate," but nonetheless fail to operate as expected. This is most clearly demonstrated by mis-specified reward functions in Reinforcement Learning, where an AI algorithm can perform very well in all the specified accuracy and performance metrics, but still behave very differently than expected.[2] The current 5.1.1 Accuracy section is most applicable to ML systems that seek to model relationships in underlying data for predictive purposes. However, this does not fully capture the training and evaluation of Reinforcement Learning systems, or systems whose outputs are recommendations or actions, rather than descriptive or predictive models about an underlying dataset.

We propose either expanding on the 5.1.1 Accuracy section, or adding a technical characteristic that better captures this risk category like the following:

5.1.1a Specification (Goal Alignment)

Specification or Goal Alignment indicates the extent to which an AI or ML system's accuracy or performance metrics are correctly aligned to its expected behavior after training. Ideally, the measured accuracy metrics for an ML system (or the objective function being optimized by a reinforcement learning algorithm) are perfectly correlated with the ultimate outcomes a human operator desires. For more complex tasks however, it is difficult to train a model using a generalized goal – such as "winning" a specific Atari game. Instead, Designers must often train algorithms against a simplified, proxy objective function.[3] This introduces risks, as an algorithm optimizing towards a proxy objective may develop behaviors that are irrelevant and even counter-productive to the ultimate objective the designer actually intended. For instance, a car racing system optimizing towards a "don't crash" proxy objective may simply drive in circles, rather than attempt to navigate the racetrack safely. To properly manage risk from misspecification, it is critical to build a testing and evaluation framework that catches these potential errors early, before impact on critical or deployed systems. Additionally, as model development moves from "toy" virtual environments to larger, real-world systems, trial and error may no longer be a sufficiently safe way of evaluating misspecification risks.

Additionally, the AI RMF defines risk simply as a function of impact and likelihood of occurrence. Getting more specific about the formulation of risk (to include threat, vulnerability, impact, and resilience) and the stakeholders for whom it matters (risk of what to whom) will make the document easier to follow, as certain sections are more abstract than others (e.g., adversarial attacks to ML systems are much more specific, as written, than interpretability risks).

Recommendation: The AI lifecycle should be expanded beyond just the AI model design, development, and test & evaluation (T&E) to include the additional activities an organization goes through to implement AI. The description of the AI lifecycle should be moved forward in the RMF document because of its importance and because it is referred to throughout the document but (currently) not shown until Section 6 (page 15).

Figure 6 on page 15 shows an AI lifecycle that is limited to just AI model design, development, and T&E. Page 15, line 3 states, "Risk management should be performed throughout the AI system life cycle," however risk management needs to be done in other phases in the AI lifecycle that are not shown in the current lifecycle diagram on page 15. The RMF would benefit from showing a more complete AI lifecycle that represents all the activities an organization must do to strategize, implement, and adopt AI. Risk is described (page 8, line 3) as inversely related to AI trustworthiness. There are things an organization can do before "pre-design" (page 15, lines 4 – 5) to address risks and assure AI trustworthiness. MITRE recommends a more representative AI lifecycle that includes four iterative phases that repeat in each evolutionary stage of an organization's AI journey.

Evolutionary stages that an organization will go through as it grows its AI experience and seeks to develop and deploy AI capabilities are:

1. Experiments
2. Proofs-of-concept and Prototypes
3. Pilots
4. Projects
5. Programs/Portfolios

The exact path or evolution from experiments to projects varies based on the AI goals, use cases, selected AI technology(ies), and other factors. Overall, the evolution progresses from "small and many" (experiments) to "large and few" (programs/portfolios). Risk and AI trustworthiness should be addressed in each evolutionary stage. Even experiments can start to consider potential opportunities and barriers related to AI trustworthiness. Rounding out the AI lifecycle with evolutionary stages and how AI risk management plays a key role in each stage will make the RMF more flexible and robust.

The four iterative phases that repeat in each evolutionary stage are:

1. Strategize: Start assessing risk and AI trustworthiness here, when the organization defines its AI strategy, selects the mission-relevant AI use cases, and determines if AI is the right technology. Business owners, ethicists, users, and other subject matter experts should be included in these strategic decisions to identify AI risks early, before the organization starts preparing for and developing AI models.
2. Prepare: Prepare the acquisitions, workforce, skillsets, cross-functional team, concepts of operations and capability needs statement, AI software platforms, compute, storage, data pipelines, data management practices, AI governance decision-making roles and responsibilities, etc. In addition to determining how to assess trustworthiness of acquired AI, the organization will also need to assess other risks associated with acquiring and operating AI solutions (such as who will monitor and maintain the AI, how to get access (if needed) to vendor code and developers and addressing vendor "AI black boxes"). The need to address these types of risk is not apparent in the current RMF.
3. Develop, Test & Evaluate, and Deploy: These are covered in the AI system lifecycle on page 15, but they are done not only for the AI system, but also for redesigned, AI-enabled business processes (workflows) that reflect human-machine teaming, new workforce roles and responsibilities, and other IT systems and interfaces that must integrate with the AI. Elements of trustworthy AI (core to the RMF) can be addressed not only in the technical AI model development, but also in human-machine teaming design, user training, workflow design, stakeholder engagement, and other activities. This way the organization builds and deploys a holistic solution, not just the "AI system."
4. Use, Monitor, and Maintain: Once AI solutions are deployed, they need to be monitored by AI/ML operations (Ops) and users. These stakeholder groups play key roles in monitoring the AI's performance against the various risk/trustworthiness factors.

The RMF focuses on technologists and the AI solution design, development, and deployment. The RMF would benefit from increased coverage of broader stakeholder involvement (including leaders, business owners, program managers, human-centered design experts, process/workflow experts, organizational change experts), and the holistic AI-enabled solution (process-people-technology) that goes beyond just the AI model design, development, and T&E.

Look beyond the component level

Recommendation: The RMF should address the fact that assessing risk at an AI model level can be different than assessing risk at a capability or system level. This issue should be addressed in Section 4,

Framing Risk, where a clear depiction of the relationship between model and system should also be made.

End-users or organizations are concerned with the risks on the capability/system (whether it is just one model or a workflow of models/components). They want to know how the system came out with the decision that it did. They don't necessarily want to know the inner workings of each model of the system. An analogy: As an end-user of a car, I don't need to know how a combustion engine works to drive a car. I just need to know that the car behaves as intended and is safe. Developers, on the other hand, are concerned with risks to the models individually as well as to the system.

For example, when discussing Accuracy in 5.1.1, the first statement is not all encompassing, *"accuracy indicates the degree to which the ML model is correctly capturing a relationship that exists within training data.*" While this is true for a single model, it may not be true for a model that is part of a system. Accuracy on a single model may be high, and risk low; however, place that model in a workflow/system and risks emerge not only from that one model, but from the evolution, interoperability needs, and emergent behaviors by the different ML-enabled components (a.k.a., models) coming together; and the resultant accuracy on a system level may be low. So, there is the need to assess risk at both the model level *and* the system level. The document should be made clearer on this issue.

(Note that ISO SAE 21434, in General Considerations, deals with this well with their definition of, and continued references to, Item and Component.)

Recommendation: The RMF should consider integration in both organizational and system context and address them in all core activities.

Section 2 Scope mentioned that AI risk "should be integrated within the organization developing and using AI and be incorporated into enterprise risk management; doing so ensures that AI will be treated along with other critical risks, yielding a more integrated outcome and resulting in organizational efficiencies." This idea is further elaborated in section 4.2.3 Organizational Integration. However, before AI risks are integrated at the organizational level, they should be assessed in the system context in which AI is a part. For example, the risk of an AI subcomponent could be mitigated by system level design.

The AI RMF should also call out how this integration could be addressed and updated within each of the core activities – map, measure, management, and govern. This recommendation relates to the system integration activities that would occur in the "Develop, T&E, and Deploy" phase of the iterative AI lifecycle, described previously.

Align and disambiguate with the existing NIST RMF

Recommendation: Greater alignment between this NIST AI RMF and the existing NIST RMF. NIST should disambiguate the titles, scope, and purpose of these two related, but different documents.

An AI-enabled system/technology inherits all the same risks as other digital systems/technologies, with only a few unique AI-specific risks. Rather than creating a totally new RMF for AI, NIST should more closely align/integrate the AI RMF into the already existing NIST RMF, SP 800-53: Prepare, Categorize, Select, Implement, Assess, Authorize, and Monitor. An AI overlay to the SP 800-53 would let the industry retain its knowledge of the current RMF and simply apply the AI-specific considerations in the context of the existing framework. This mapping will require an analysis of existing controls that will need to be modified and or new controls to be created.

This approach will avoid confusion by matching the scope of the existing RMF process. These changes in the AI RMF document's scope would enable it to be more rapidly adapted into practice. For example, this would greatly aid the Department of Defense (DoD) as they have successfully implemented its derivative from NIST RMF, SP 800-53, establishing a framework and controls for protecting DOD systems from cybersecurity threats.

The MITRE support team supporting Office of the Secretary of Defense (OSD), Developmental Test, Evaluation, and Assessments (DTE&A) is developing a definition of the future state of T&E of AI. This product was shared at a recent Workshop, hosted by DTE&A. Workshop recommendations included maturation and adoption of a RMF to address cybersecurity challenges to AI, complementary to the approach being taken by NIST.[4]

Greater emphasis on risks arising from Adversarial ML

Recommendation: Separate the Resilience and Security characteristics and expand discussion on each including the unique issues stemming from the use of AI technologies – particularly machine learning.

The AI RMF groups Resilience and ML Security at every occurrence in the document for example as a singular technical characteristic in Section 5.1.4. This combination mixes two separate but interrelated characteristics in a way that limits the level of focus and enumeration needed within the AI RMF if it is to empower and enable the identification and mitigation of unique risk associated with AI employment.

NISTIR 8269[5] (Draft) defines resilience as the ability to withstand, adapt to, and recover from adverse conditions. The conditions may or may not arise from adversarial actions. Section 5.1.4 should be split into two parts. Resiliency is an important characteristic, and it should have its own section discussing its meaning and the types of conditions a system might be resilient to.

A section on the technical characteristic of Security focused on the implications of Adversarial ML deserves full discussion. Resources such as NISTIR 8269 and the MITRE Adversarial Threat Landscape for Artificial-Intelligence Systems (ATLAS) Matrix[6] describe both attack paths and consequences, showing how the risks resulting from adversarial actions are well-known and significant. Examples of such attacks include an evasion attack on a computer vision system, where for example, a sticker on or near a banana causes it to be misclassified as a toaster.[7] The implication for risk of object misclassification is significant in a critical system such as autonomous vehicle vision.

As an example, these technologies are being aggressively adopted for authentication as well as identity verification systems. Implementations of this technology range from Apple's Face ID to unlock an electronic device, to walking through a passport screening system at an international airport. Unfortunately, as face-based identification and authentication systems become more ubiquitous, attempts to defeat them have also risen.

ID.me reported that between June 2020 and January 2021 it identified more than 80,000 attempts to fool their verification system in attempts to fraudulently access state financial benefits.[8] Would be attackers used a broad range of techniques ranging from the extremely simple, such as wearing a picture of a person on their face, to highly sophisticated attacks leveraging adversarial AI. It is difficult to assess the true impact of these attacks as we are only able to see the attacks which were unsuccessful, but analysts at Experian PLC see this as a part of a fast-growing type of financial crime called synthetic identity fraud.[9] Most of the would-be attackers were the equivalent of cyber "script kiddies" who used off the shelf AI tools such as "this-person-does-not-exist.com" to generate easily detected AI generated faces or various printed or homemade masks. However, in an example of what a more determined and sophisticated adversary can accomplish, the Shanghai Hongkou District People's Procuratorate prosecuted Wu Moumou and Zhou Mouof for stealing US$76 million by exploiting AI vulnerabilities in the government's facial recognition service.[10]

In another case, Microsoft researchers from China demonstrated the broad susceptibility of facial recognition-based authentication to the injection of back doors directly into the AI models.[11] The researchers began by scraping the Android Appstore for applications which leverage deep learning models within their code base. Out of the 116 identified, they were able to successfully inject back doors into 54 of the applications. While the researchers did not list the application names out of the 54 vulnerable applications, five were financial services applications which together represented over 120,000,000 downloads.

As these AI-based attacks continue to grow, efforts such as the NIST AI RMF must provide the conceptual tools necessary to empower private and public sector organizations to address the challenges and complexities inherent in securing AI-enabled systems against adversarial attack. This is of particular importance as many of these attack vectors fall through the cracks between current cybersecurity and risk frameworks. Separating ML Security from Resilience will enable highlighting the unique challenges faced due to adversarial attacks as well as drive the necessary focus and thought leadership to ensure the NSF AI RMF adequality addresses these challenges.

Additionally, the companion practice guide should enumerate how the AI RMF can be used along with industry AI security efforts such as the Partnership on AI's Incident Database and MITRE ATLAS as well as highlight available AI vulnerability assessment tools such as Microsoft's Counterfit[12] and IBM's Adversarial Robustness Toolbox.[13] Providing guidance on how industry can leverage these capabilities to implement the AI RMF in real world scenarios will be critical to ensure the applicability and usability of the AI RMF.

II. Other Issues

Recommendation: NIST should rethink how "users" are represented across the key stakeholder groups defined in Figure 1 to better support certain types of policy implementation for AI risk management.

There is at least one fundamental and potentially fatal breakdown between the stakeholder groups defined in this document (as depicted on page 4 in Figure 1). In real world deployments, there is a spectrum of "user" or "operator" that transcends the current Operators & Evaluators Stakeholders group and the General Public Stakeholders group. One question up front is what is the difference between a "user" and an "operator"? If such a distinction is necessary to manage the risks of an AI system, then the difference should be made clear in the figure and supporting text. Assuming there is no need to make such a distinction, then we are left with a spectrum of skilled user – from expert to non-expert. Expert users are currently associated with the Operators & Evaluators Stakeholders group, but non-expert users are left to be associated with "individuals" and "consumers" in the General Public Stakeholders group. This current stakeholder group delineation is problematic for setting policies generally applicable to all "users", for example, when trying to identify and associate responsibilities for AI system accountability later in Section 5.3.2. Should only expert users be held accountable for proper use of an AI system? Or should all users including non-expert consumers be held to some level of accountability? This fragmented handling of "users" in the stakeholder groupings creates in some policy contexts a false dichotomy.

In addition, should AI users be held accountable for AI system performance? The current RMF states that users should be accountable, when foremost it is AI developers, leaders, and their organizations who are accountable for performance of the AI systems deployed. The users must be involved up-front to identify meaningful, feasible AI use cases, provide human-centered design input, and work with the cross-functional AI team to define the new AI-enabled work processes. When the AI is developed, users should be involved iteratively in test, evaluation, validation, and verification (TEVV) of the AI solution. Before the AI deploys, users must receive thorough communication and training on how they will use the AI—what outputs of the AI will people use to do what specific day-to-day work tasks. Users may be accountable for performance depending on the level of human-machine interaction required in the AI-enabled work process. AI can aid, augment, or automate a human task. The degree of human control, directability, and interaction with the AI are among the many factors that would determine level of user accountability for AI system performance.

Recommendation: The definition of risks, the categories, and the distinction between risks and characteristics, should be clarified.

Any AI RMF should help the stakeholder groups of Figure 1 identify specific risks that can then be tracked and addressed, either by correcting the underlying issue or accepting the risk as part of a governance process. It is observed that Figure 2 lists examples of potential harms from AI systems which is one way to frame types of risk, but this does not cover all the categories of risk that must be managed for AI systems. Figure 3 is mislabeled "Risks & Characteristics" containing only the *characteristics* of a trustworthy AI system, which are not risks in and of themselves. The AI RMF

should make it more clear how harms and trustworthy AI characteristics are used to define and identify actual risks within an AI system.

Recommendation: NIST should provide examples of different challenges facing small to medium-sized organizations (over large organizations) when implementing the AI RMF.

On page 7, lines 32-33, it would be illuminating to the reader if NIST were to provide some examples of different challenges facing small to medium-sized organizations (over large organizations) when implementing the AI RMF. This will help smaller organizations understand, prepare, and resource such endeavors for success.

Recommendation: "Traceability" should be added to the taxonomy of trustworthy AI characteristics.

Traceability, having policies and procedures in place to ensure that models and underlying code are traceable and auditable, is an important characteristic of AI solutions. Those who develop AI versus those who deploy it may be from entirely different organizations within a larger structure or completely independent (vendor). Those who deploy AI are typically responsible for policies and procedures. The deployers need to interact with the model developers in order to understand its strength and limitations. Vendors and/or contractors who develop an AI solution may no longer be around when it is time to perform ongoing monitoring. If the organization's AI/ML Operations group detects drift or other performance variances, the model may need to be retrained, redeployed, or retired. If the model was developed by a vendor or contractor, and/or if the organization has no knowledge of traceability of the code or model(s), then this presents a serious risk to sustaining operation of the AI solution. The term "Traceability," should be added to Figure 3 on page 8 under Technical Characteristics, it should be added into Figure 4, and it should be described in a sub-section within Section 5.1.

Recommendation: Discuss how risk management fits with AI governance, which includes decision-making and communication, assuring AI trustworthiness, and oversight of AI efforts.

Page 14, line 14 states, "Assuming a governance structure is in place…". This is a very big assumption. Many organizations lack an AI governance structure. Having a governance structure, even if not elaborate or complex, is essential for effective AI risk management. Risk management, as described in the RMF, should be deliberately addressed by the organization in a holistic manner, either as part of its AI Governance activities or its AI Project & Risk Management activities. A holistic, multi-disciplinary approach is necessary to achieve the ultimate goal of successful AI adoption. Having AI governance in place means: there are defined roles and responsibilities for making and communicating decisions about the organization's AI endeavors; there is a coordinated, organized way for assuring trustworthiness of AI solutions (which the RMF focuses on); and there is ongoing oversight of AI efforts (experiments, proofs-of-concept, prototypes, pilots, projects, programs) and learning from these efforts to inform the organization's next iteration/refinement of its AI strategy.

Additionally, MITRE's response to NIST Study to Advance a More Productive Tech Economy – AI (Document No. 21116-0234) provided MITRE recommendations on AI governance and things to include in the risk framework.

Recommendation: The document should do more to address what policy makers should do to assess AI risks.

The acceleration of AI technology, along with the exponential growth of available data, has dramatically increased the demand for better AI risk mitigation policies in government, academia, and the commercial sector. Policies and infrastructure for AI risk need to be consonant with technology development and modernization, particularly in the context of AI T&E, breaches in data security, AI/ML attacks, and privacy preservation. Organizations need to develop and implement AI risk management policies that: define and guide their workforce in the missions they support; detail how much risk should be or can be assumed; describe the mission impact of identified risks; and delineate corresponding leadership roles and responsibilities pursuant to assumption of these risks. This reiterates the importance of having sound AI governance and risk management activity threads with clear decision-making roles and responsibilities, as described above.

Recommendation: References to "organization" across the AI RMF should be qualified within the context of use as the stakeholder landscape covers a variety of organizations.

For example, the reference to "organization" on page 14 in lines 22-23 should be qualified. Is the text calling out the organization that designed and developed the AI? The organization that purchased and deployed the AI? The organization that is using the AI? The responsibilities of the AI RMF should be carried out in some respects by all these various organizations. This leads to the problematic use of the qualifier, "potentially" within this section. In general, External Stakeholders and General Public Stakeholders are in scope to the RMF as called out in Figure 1, so the qualifier "potentially" should be struck, and the paragraph should be rewritten to appropriately address all relevant organizations.

Recommendation: The term "override" should either be further discussed in this document or omitted.

The term "override" is mentioned two times in this document (on page 15, line 7, and on page 18 in Table 3) without definition or discussion. If this element of deployment is going to be included in the AI RMF, it should be more fully addressed. Particularly, because override is not always an available design feature depending on the use case, and because providing a mechanism for overriding an AI system may in itself introduce unintended risks. It is recommended that the term "override" be omitted and instead the framework continue to focus on the need for actionable redress.

Recommendation: Expand the incorporation of "actionable redress" within Manage and Govern functions.

In the entire draft AI RMF, "redress" is only mentioned in passing twice on page 13. With the potential performance uncertainties of AI systems – whether due to changes in the operational evironment (i.e.,

drift), unexpected societal outcomes, etc. – the AI RMF should incorporate mechanisms for actional redress to protect potentially impacted individuals and communities from harm caused by incorrect or adverse AI system outputs. Organizations deploying AI systems may need to be more proactive with redress – and reach out to wrongly impacted individuals who may not be aware of the AI incorrect result. Redress should be included as a Manage function category/subcategory in the context of the interative AI adoption lifecycle phase – Use, Monitor, and Maintain – described above.

The protections afforded by actionable redress should also be incorporated within the Govern function. The following text is recommended for a new subcategory under Table 4, Category 5, on page 19 – "Processes are in place to support actionable redress related to incorrect or adverse AI system outputs."

Recommendation: The Govern function should include stakeholders "using" the AI system.

For example, on page 18, line 3, why would the Govern function not also include "using" AI systems? Right now, the list if limited to developing, deploying, or acquiring. The document seems to make it clear that risks are also introduced with the "use" of an AI system. Individuals involved in using AI systems should also be included with the list on page 18, line 8.

Recommendation: The 5.2.3 Privacy section should include legal considerations, for example involving the examination of human data.

Stringent data protection practices need to be in place within an organization that develops AI systems to effectively secure personally identifiable information (PII), rigorously protect training data and models, and prevent adversarial attacks. Consideration of the legal dimension attendant to the proper protections and use of data and AI models should be discussed and emphasized in this section on Privacy.

Recommendation: The assessment of whether the AI is the right tool to solve the given problem (e.g., if the system should be further developed or deployed) should be performed earlier. This assessment should take place at the time of pre-design before the initial Map function begins.

Whether AI is the right tool to solve the given problem is an important decision step but is currently hidden within a subcategory of the Manage function. Determination factors for whether AI is the right tool include business value, cultural readiness, and availability of training data. These factors impact all the AI RMF core activities from mapping to governing. Therefore, the assessment should be performed earlier at the time of pre-design before the initial Map function begins. The Strategize phase recommended in MITRE's more holistic AI lifecycle (above) addresses the need to determine up-front the key mission problems, whether AI is the right technology to solve these problems, and who the business owner is for the AI solution. Determining AI fit and feasibility early is one of the reasons it is important for an organization to deliberately go through the Strategize and Prepare phases before beginning AI development.

Recommendation: Section 5 should include and make clear the role the general public may play as impacted stakeholders.

The 'general public' is defined on page 5 as being 'most likely to directly experience positive and adverse impacts of AI technologies.' However, 'general public' is not mentioned in Section 5 – and as impacted stakeholders, they have a role to play. 'Impacted individuals' and 'impacted communities' are mentioned once each in Section 5. Impacted individuals/communities should be partners to the extent feasible - and have more of a role throughout the AI RMF.

Recommendation: The potential impact associated with an emerging AI solution should be a starting point for analysis and influence just about every category/subcategory in terms of identifying the need for independent review and analysis.

The concept of independent review is mentioned in the 4th subcategory of the 4th category for Map but should be more pervasive throughout the RMF given the amount of uncertainty with AI.

III. Line Edits

Item	Page	Line	Description / Comment	Original Text	Suggested Text
1	1	34	Because AI risk management (and addressing trustworthy AI) spans activities upstream (prior to) development, deployment, and use (see the AI lifecycle comments above), MITRE recommends that line 34 be reworded	the development, deployment, and use of AI systems.	the strategizing, planning, preparation, design, development, evaluation, deployment, use, and monitoring of AI systems.
2	2	26-28	Delete "organized and" and "understood and" to improve clarity of this sentence.		
3	3	11		between	among
4	3	15-17	Attribute 5 includes numerous goals (be "usable," "mesh with other aspects of risk management," be "intuitive," be "readily adaptable as part of an organization's broader risk management strategy and processes," be "consistent or aligned with other approaches to managing risks"). Note that "other approaches to managing AI risks is ambiguous – if there are specific AI risk managing approaches that the authors envision as being essential for the AI RMF to "mesh" with, it is important to name them. I suggest streamlining this attribute to focus on a single goal:		Be readily adaptable by using approaches consistent with broader risk management strategies that organizations currently use.
5	3	19	"technology agnostic" It is not clear what this phrase means in the context of the AI RMF. This term is not plain language. Suggest defining it. Also, if it is essential to keep, suggest that it be hyphenated ("technology-agnostic") (in keeping with Attribute 9 ("law- and regulation-agnostic").	technology agnostic	technology-agnostic
6	3	22	Delete "Take advantage of and"	Take advantage of and foster greater awareness	Foster greater awareness
7	3	25	Suggest removing "Be law- and regulation-agnostic." This phrasing is not plain language. The second sentence is clear and		

			more accurate, although broad: "The AI RMF is meant to be support organizations' abilities to operate under applicable domestic and international legal or regulatory regimes."		
8	3	27	Delete "readily"		
9	3	31	Delete "relatively".		
10	4	3	These are four stakeholder categories, rather than groups. The items within each category are the stakeholder groups.	groups	categories
11	4		In Figure 1 and Line 12: Change "business teams" to "business owners". These are the stakeholders who ultimately have to buy-in to using AI in their mission/business operations.	business teams	business owners
12	4	8-10		AI system stakeholders are those who have the most control and responsibility over the design, development, deployment, and acquisition of AI systems, and the implementation of AI risk management practices.	AI solution stakeholders are those who have the most control and responsibility over the strategizing, planning, acquisition, design, development, deployment, and monitoring of AI systems, process/workflow changes, people and organization changes, and the implementation of AI risk management practices.
13	4	10-13		They may include individuals or teams within or among organizations with responsibilities to commission, fund, procure, develop, or deploy an AI system: business teams, design and development teams, internal risk management teams, and compliance teams.	They may include individuals or teams within or among organizations with responsibilities to decide, select, commission, fund, procure, plan, prepare, design, develop, deploy, or use an AI system and associated changes to policies, processes/workflows, role and responsibilities, workforce skills, data management practices, IT infrastructure and interfaces. These individuals and teams include senior leaders, business owners, program/portfolio managers, acquisition managers, data managers, human-centered design experts, process owners, organizational change management experts, training experts, ethicists, community representatives, AI design and development

					teams, internal risk management teams, and compliance teams.
14	4	20-21 and Figure 1	Researchers. Suggest that Researchers be included in the inner circle ("AI System") as well as the "Operator and Evaluator" circle. It is not possible to have development and designs teams with researchers involved. In other words, the core circle must include researchers since it is sphere in which stakeholders have "the most control and responsibility" – it would be irresponsible to put researchers (SMEs) only in operator and evaluator roles. Expertise must inform design, development, deployment, and acquisition of AI systems.		
15	5	5	Unclear what is meant by "They [the general public] may provide the motivation for actions taken by other stakeholders […]". Suggest revising this sentence. Note that as written, the focus is on members of the public directly affected by AI systems ("individuals, communities, and consumers in the context where an AI system is developed or deployed"). This is both vague and narrow. Suggest stating more clearly that the concern with preventing or mitigating harmful impacts (direct or indirect, short- or long-term) to individuals and groups.		
16	5	19-22	This definition of risk does not include vulnerability or resilience. I suggest adding these elements of risk to the definition – unclear where the current definition comes from or why it is authoritative in this context (add references). A robust risk definition is missing from section 4.1 – without this, later sections such as "Risk Measurement" lack clarity (what must be measured, how, and why only makes sense in light of an risk definition).		

17	6	2	Missing period after "applications."	applications	applications.
18	6	14	Suggestion	latent at present but may increase in the long term as AI systems evolve	latent at present but may increase in the long term as AI systems and their use evolve
19	6	19	Section 4.2.2 is entitled "Risk Thresholds" and yet it speaks more broadly to risk factors, along with their thresholds and values. The section would be more appropriately entitled, "Risk Factors, Thresholds, and Values". Note that the document at times uses the terms risk factors, characteristics, and attributes interchangeably. Characteristics and factors are used in earlier sections, whereas Section 6 uses attributes. It is suggested that NIST authors choose one term and use it consistently throughout the document.	"Risk Thresholds"	"Risk Factors, Thresholds, and Values"
20	6	20-23	Key Risk Indicators are not limited to just "thresholds", but indicators are rather factors along with associated thresholds or values. Suggest rewriting these sentences to more accurate reflect this.	Thresholds refer to the values used to establish concrete decision points and operational limits that trigger a response, action, or escalation. AI risk thresholds (sometimes referred to as Key Risk Indicators) can involve both technical factors (such as error rates for determining bias) and human values (such as social or legal norms for appropriate levels of transparency).	Key risk indicators include factors along with associated thresholds or values. Factors can be technical (such as error rates for determining bias) and they can be human values (such as social or legal norms for appropriate levels of transparency). Thresholds refer to the values used to establish concreate decision points and operational limits that trigger a response, action, or escalation.
21	7	7		The AI RMF does not prescribe risk thresholds or values.	The AI RMF does not prescribe specific risk factors nor their associated thresholds or values.
22	7	10	It is important to include business owners here, and generally in more places throughout the RMF. The RMF focuses on technical stakeholders and often leaves out the leaders and business owners to whose mission operations the AI is supposed to deliver benefit and value.	risk thresholds should be set through policies and norms that can be established by AI system owners, organizations,	risk thresholds should be set through policies and norms that can be established by leaders, business owners, AI system owners, organizations,
23	7	11		Risk thresholds and values are likely to	Risk factors, thresholds, and values are likely to
24	7	20-21		The AI RMF provides the opportunity for organizations to specifically define their risk	The AI RMF provides the opportunity for organizations to specifically define their risk

				thresholds and then to manage those risks within their tolerances.	factors and thresholds to manage those risks within their tolerances.
25	7	29-31	Recommended edits	Effective risk management needs organizational commitment at senior levels and may require significant cultural change for an organization or industry.	Effective risk management needs effective organizational change management, which includes stakeholder engagement, communication, and organizational commitment at senior levels and may require significant cultural change for an organization or industry. Culture is the most difficult thing to change in an organization. Therefore, it is imperative that organizations involve the right people and address AI fit, feasibility, and trustworthiness concerns up-front when the organization is defining its AI strategy, selecting AI use cases, and choosing AI efforts to pursue.
26	8	Figure 3	Hyphenate Socio-Technical	Socio Technical	Socio-Technical
27	8	17-28	The use of the terms convergent validity and discriminant validity are not appropriately applied in a helpful manner to the reader. Tying these to "data" is ambiguous – whether the author is referring to ML training data or the later mentioned experimental data. Regardless, validity in context here is being applied firstly to the AI system, whether experimentally or operationally. Also, there is no reason to limit assessment of AI validity to primarily ML models.	Technical characteristics in the AI RMF taxonomy refer to factors that are under the direct control of AI system designers and developers, and which may be measured using standard evaluation criteria. Technical characteristics include the tradeoff between convergent-discriminant validity (whether the data reflects what the user intends to measure and not other things) and statistical reliability (whether the data may be subject to high levels of statistical noise and measurement bias). Validity of AI, especially machine learning (ML) models, can be assessed using technical characteristics. Validity for deployed AI systems is often assessed with ongoing audits or monitoring that confirm that a system behaves as intended. It may be possible to utilize and automate explicit measures based on variations of standard statistical or ML techniques and specify thresholds in requirements. Data generated from experiments that are designed to evaluate system performance also fall into	Technical characteristics in the AI RMF taxonomy refer to factors that are under the direct control of AI system designers and developers, and which may be measured using standard evaluation criteria. Validity of AI can be assessed using technical characteristics. These characteristics can be evaluated for convergent validity, discriminant validity, and statistical reliability. It may be possible to utilize and automate explicit measures based on variations of standard statistical techniques and specify thresholds in requirements. Validity for deployed AI systems is often assessed with ongoing audits or monitoring that confirm that a system behaves as intended. System performance may also be evaluated experimentally including tests of causal hypotheses and assessments of robustness to adversarial attack.

				this category and might include tests of causal hypotheses and assessments of robustness to adversarial attack.	
28	9	4	Suggest removing "(or ML Security)" and leaving the discussion of ML security to Section 5.1.4. This is not meant to diminish the importance of AI security, but to not overly constrain the risk attribute of resilience and unnecessarily limit the application of this document to ML-based AI systems.	reliability, robustness, and resilience (or ML security).	reliability, robustness, resilience and security.
29	9	6		Accuracy indicates the degree to which the ML model is correctly capturing a relationship that	Accuracy, in the case of an ML-based system, indicates the degree to which the AI model correctly captures a relationship that
30	9	8	The metrics listed are not specific to ML.	standard ML metrics	standards performance metrics
31	10	1		ML datasets or models in ways that cause	training datasets or models in ways that cause
32	10	4-17	The definitions of Robustness and Resilience do not align well with the definitions provided in the Draft NISTIR 8269.		(See NISTIR 8269 for definitions)
33	10	13-17		A model that can withstand adversarial attacks, or more generally, unexpected changes in its environment or use, may be said to be resilient or secure. This attribute has some relationship to robustness except that it goes beyond the provenance of the data to encompass unexpected or adversarial use of the model or data. Other common ML security concerns relate to the exfiltration of models, training data, or other intellectual property through AI system endpoints	In general, resilience is an AI model's ability to withstand changes in its environment or use, and more specifically, its ability to withstand adversarial attack. This attribute has some relationship to robustness except that it goes beyond the provenance of the data to encompass unexpected or adversarial use of the model or data. In the case of ML-based systems, security concerns include exfiltration of models, training data, or other intellectual property through AI system endpoints.
34	11	13		work to users' skill levels	work articulated to the skill level of the user
35	11	17		the user lacks an understanding of ML technical principles	the user lacks an understanding of underlying technical principles
36	11	19	Delete "Interpretability seeks to fill a meaning deficit." This is not plain language. The definition of interpretability follows in the next sentence.		

37	11	22	Delete "The underlying assumption is that" –Even if this clause is removed, the meaning is unclear, as perceptions of risk stem not only from lack of ability to appropriately make sense of model outputs. In fact, the key point seems to be that perceptions of risk are contingent upon making sense of model outputs, a process that may be negatively impacted if models are not interpretable or, if interpretable, are interpreted incorrectly.		
38	11	18-31	This section is difficult to understand. Suggest rewriting in plain language. The authors should provide evidence for this claim.		
39	12	19-26	The lists in this callout box that exemplify 3 cases are difficult to parse and distinguish based on the use of commas alone. Suggest the following revision.	When managing risks in AI systems it is important to understand that the attributes of the AI RMF risk taxonomy are interrelated. Highly secure but unfair systems, accurate but opaque and uninterpretable systems, and inaccurate, but fair, secure, privacy-protected, and transparent systems are all undesirable. It is possible for trustworthy AI systems to achieve a high degree of risk control while retaining a high level of performance quality. Achieving this difficult goal requires a comprehensive approach to risk management, with tradeoffs among the technical and socio-technical characteristics.	When managing risks in AI systems it is important to understand that the attributes of the AI RMF risk taxonomy are interrelated. For example, the following are all undesirable: a. highly secure but unfair systems; b. accurate but opaque and uninterpretable systems; and c. inaccurate, but fair, secure, privacy-protected, and transparent systems. It is possible for trustworthy AI systems to achieve a high degree of risk control while retaining a high level of performance equality; however, achieving this challenging goal requires a comprehensive approach to risk management, often with tradeoffs among technical and socio-technical characteristics.
40	12	30	The authors should cite the OECD Principles on Artificial Intelligence, which would provide a stronger basis than appeal to general agreement (line 30: "it is widely agreed").		
41	12 And 13	35-38 1-3	This paragraph is problematic in that it attempts to provide overly abbreviated meanings of fairness, accountability, and transparency which become inadequate and		Guiding principles that are relevant for AI risk include fairness, accountability, and transparency.

			redundant to the more detailed treatments that immediately follow in the subsections.		
42	12	37	Even if the human operators (users) of AI solutions are in the organization, it is unclear how the operators/users would be accountable for the AI's outcomes. The leaders, decision-makers, and developers would be accountable. The phrase "and their organizations" takes care of leaders and decision-makers. This section also gets to the risks an organization assumes if it acquires an AI solution from a vendor: who is accountable for acquired AI. Managing risks of acquired AI is another part of AI risk management that needs more coverage in the RMF.	Individual human operators and their organizations	Individual developers and their organizations
43	13	5-13	The good statement tying issues of equity, bias and discrimination cut from the preceding introductory paragraph should be moved to this subsection. Plus, the statement related to process fairness is best presented as past tense. The following treatment of fairness and the existence of harmful systems is awkward and would benefit from a rewrite.	Standards of fairness can be complex and difficult to define because perceptions of fairness differ among cultures. For one type of fairness, process fairness, AI developers assume that ML algorithms are inherently fair because the same procedure applies regardless of user. However, this perception has eroded recently as awareness of biased algorithms and biased datasets has increased. Fairness is increasingly related to the existence of a harmful system, i.e., even if demographic parity and other fairness measures are satisfied, sometimes the harm of a system is in its existence. While there are many technical definitions for fairness, determinations of fairness are not generally just a technical exercise. Absence of harmful bias is a necessary condition for fairness.	Fairness in AI systems includes concerns for equality and equity by addressing socio-technical issues such as bias and discrimination. Standards of fairness can be complex and difficult to define because perceptions of fairness differ among cultures. For one type of fairness, process fairness, AI developers had made a practice of assuming ML algorithms were inherently fair because the same model-building processes were applied regardless of the user. This perception has eroded as awareness of biased AI algorithms and biased datasets have been observed and their harms documented. While there are many technical definitions for fairness, determinations of fairness are not generally just a technical exercise. Even if demographic parity and other fairness measures are satisfied, it is still possible for an AI system to cause harms. Nonetheless, absence of harmful bias is a necessary condition for fairness.
44	13	15-21	The lead-in sentence to section 5.3.2 is extremely awkward. Also, the attribution of	Determinations of accountability in the AI context are related to expectations for the	AI systems should be designed and deployed in such a way that AI systems along with

			who is responsible and accountable in this paragraph is inappropriately limited and ascribed. In principle, aspects of accountability should be assigned to AI system designers, developers, evaluators, vendors, and users (whether individual or organizations). This discussion of who is accountable is trending toward how to ascribe liability, which is a divergence in topic from that of managing risk and should be avoided in this document. The last sentence of the paragraph, "Grounding organizational practices …" doesn't make sense and doesn't add anything meaningful. Suggest rewriting the paragraph as follows. Note that if this use of the Figure 1 stakeholder groups is not appropriate for assigning accountability, then it is a sign that these stakeholder categories are flawed and not sufficient for managing AI risks.	responsible party in the event that a risky outcome is realized. Individual human operators and their organizations should be answerable and held accountable for the outcomes of AI systems, particularly adverse impacts stemming from risks. The relationship between risk and accountability associated with AI and technological systems more broadly differs across cultural, legal, sectoral, and societal contexts. Grounding organizational practices and governing structures for harm reduction, like risk management, can help lead to more accountable systems.	their associated stakeholders (i.e., AI system, operators, and evaluators stakeholders in Figure 1) are held accountable while protecting external and general public stakeholders from adverse impacts stemming from risks. The relationship between risk and accountability associated with AI (and technological systems more broadly) differs across cultural, legal, sectoral, and societal contexts. Accountability is necessary for directing actionable redress related to incorrect and adverse AI system outputs.
45	13	16	Suggest replacing the term "risky" with "adverse" (since "adverse" is used throughout the document).	risky	adverse
46	13	22-32	Note that the definition of "transparency" focuses exclusively on "a user," whereas elsewhere in the document other stakeholders (oversight, decision-makers, etc.) are included. Suggest making transparency broader so it is not just about user who directly interact with the system, but about the many other stakeholders who are informed or impacted by it.		
47	13	23-24	The information imbalance referenced here is more accurately between AI system designers / developers and AI system users. Designers and developers are associated in Figure 1 to the AI Systems Stakeholder group, not the Operators & Evaluators Stakeholder group. The NIST writing team should make sure that all references to		

			parties within the AI RMF stakeholder landscape are consistently tied to the categories in Figure 1.		
48	13	32	"Desiderata" is not plain language. Suggest replacing the term.		
49	16	6-7	A new term "event" is introduced without qualification or definition.	An event can have multiple causes and consequences and can affect multiple objectives.	An adverse event can have multiple causes and consequences and can affect multiple risk managing objectives.
50	17	Table2, ID 1	Suggest replacing "elicited" with "mapped" making it clear these system requirements are being obtained from the Map function.	Elicited system requirements are analyzed.	Mapped system requirements are analyzed.
51	17	Table2, ID 2	ML security is a subset of resilience.	resilience (or ML security)	resilience (including ML security)
52	17	Table2, ID 2	Rephrase second subcategory as the end of the sentence is unnecessarily redundant when it comes to available measurement techniques.	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.	Mechanisms for tracking identified risks over time are in place, particularly if potential risks are difficult to assess using available measurement techniques.
53	18	Table3, ID 2 3rd sub-category	Should add additional qualification of harm to stakeholders.	Mechanisms are in place and maintained to supersede, disengage, or deactivate existing applications of AI that demonstrate performance or outcomes that are inconsistent with their intended use.	Mechanisms are in place and maintained to supersede, disengage, or deactivate existing applications of AI that demonstrate performance or outcomes that are inconsistent with their intended use or otherwise cause harm to stakeholders.
54	19	Table4 ID 2, 1st sub-category	Responsibilities and accountability to AI risks extend beyond the nebulous "organization" and should be clear and transparent to across the entire internal and external stakeholder landscape. Suggest the following revision.	Roles and responsibilities and lines of communication related to identifying and addressing AI risks are clear to individuals and teams throughout the organization.	Roles and responsibilities and lines of communication related to identifying and addressing AI risks are clear to individuals, teams, and groups across the entire internal and external stakeholder landscape.
55	19	Table4 ID 2, 3rd sub-category	Given the diverse stakeholder landscape of various groups and organizations, the particular organization being addressed here should be clearly qualified. Suggest the following revision.	Executive leadership of the organization considers decision about AI system development and deployment ultimately to be their responsibility.	Executive leadership of organizations responsible for development and deployment of AI systems should fully accept responsibility for AI risk management.

IV. Endnotes

[1] DARPA for example is strategically investing in driving new AI capabilities toward the next generation of AI. See DARPA's AI Next Campaign at https://www.darpa.mil/work-with-us/ai-next-campaign.

[2] Faulty Reward Functions in the Wild (openai.com): https://openai.com/blog/faulty-reward-functions/

[3] Key Concepts in AI Safety: Specification in Machine Learning - Center for Security and Emerging Technology (georgetown.edu), https://cset.georgetown.edu/publication/key-concepts-in-ai-safety-specification-in-machine-learning/

[4] "Future State of Test & Evaluation of Artificial Intelligence," MITRE, April 13, 2022.

[5] https://csrc.nist.gov/publications/detail/nistir/8269/draft

[6] https://atlas.mitre.org/

[7] Tom B. Brown, Dandelion Mané, Aurko Roy, Martín Abadi, Justin Gilmer. "Adversarial Patch," 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA. (https://arxiv.org/pdf/1712.09665.pdf)

[8] https://cacm.acm.org/news/253981-faces-are-the-next-target-for-fraudsters/fulltext

[9] https://www.experian.com/decision-analytics/synthetic-identity-fraud

[10] South China Morning Post, "Chinese government-run facial recognition system hacked by tax fraudsters: report" https://sg.news.yahoo.com/chinese-government-run-facial-recognition-102910731.html

[11] Li, Y., Hua, J., Wang, H., Chen, C., Liu, Y. "DeepPayload: Black-box Backdoor Attack on Deep Learning Models through Neural Payload Injection." arXiv:2101.06896v1, Jan 2021.

[12] https://www.microsoft.com/security/blog/2021/05/03/ai-security-risk-assessment-using-counterfit/

[13] https://adversarial-robustness-toolbox.org/