Carnegie Mellon University
Software Engineering Institute

COMMENTS ON NIST AI RMF RFI

Rachel Dzombak, Ramayya Krishnan, Carol Smith, Brett Tucker, and Nathan VanHoudnos

September 2021

NIST is soliciting input from all interested stakeholders, seeking to understand how organizations and individuals involved with designing, developing, using, or evaluating AI systems might be better able to address the full scope of AI risk and how a framework for managing AI risks might be constructed.

We present our comments and suggested changes in prose format, in addition to the provided Excel spreadsheet, for readability.

1. The greatest challenges in improving how AI actors manage AI- related risks—where ‘‘manage’’ means identify, assess, prioritize, respond to, or communicate those risks;

Comment:

In traditional systems engineering and risk, we have a model of the system to which we can apply statistical and decision theoretic approaches to risk management. With AI Systems, both the system structure and system state are evolving, and the time constants on the dynamics of systems state and systems structure are different. All of that contributes to the complexity of AI systems.

One of the greatest challenges is getting actors to see the whole system and hold the inherent complexity. Many want to approach AI systems and their risks linearly, tracking cause and effect. With AI, a necessary shift is to consider emergent issues and risks as components of interconnected and interacting systems rather than as independent issues with unrelated consequences. Addressing a risk likely means creating new vulnerabilities and new systems tradeoffs. Improvements in management of AI-related risks requires new approaches that reflect a whole systems perspective. As part of that, organizations need new approaches that broaden the scope of risk-based decisions to include opportunistic risk as well as possible threats.

More generally, an AI system can only address risks that are known and within the purview of the system. An additional great challenge in improving the management of risk is then in making systems that can deal with complexity and which are built in ways that consider broad sets of risks. This level of risk “management” will not be possible in the near term (5-10 years?) without human oversight. Furthermore, as the system evolves over time, the system will exhibit emergent behavior (e.g., a new link is learnt between two concept nodes in a graph structure). Work is needed to understand what processes and human behaviors can assess structure changes and corresponding risks before they are accepted.

In the short term, as systems are trained to anticipate and recognize risks, the challenge will be communicating both known and potential emergent risks to the people that are working with the system

in an appropriate and clear way – the design of the end-user's experience in this regard can be significant.

Suggested change:

First, consider including response options of "enhance, exploit, and share" that can go along with the original response options provided that include, "avoid, mitigate, transfer, and accept". This will allow organizations to strike a balance between possible threats and opportunities. Currently, the question is posed for threats only.

Second, consider including a direct focus on end user experience.

2. How organizations currently define and manage characteristics of AI trustworthiness and whether there are important characteristics which should be considered in the Framework besides: Accuracy, explainability and interpretability, reliability, privacy, robustness, safety, security (resilience), and mitigation of harmful bias, or harmful outcomes from misuse of the AI;

Comment:

There are different kinds of trust that arise in AI systems. There is the trust that a user places on the recommendations of an AI system in the human-AI setting (be it dyadic or team) which is determined as much by how robust the recommendations are, understandability and error rates (type 1 and 2) and the costs of errors and who bears it (think lawsuits and insurance). Another kind of trust is the community or individuals who are affected by the actions taken via the AI system and their assessment of whether the systems is trustworthy (think police and its use AI systems for face recognition).

These characteristics of AI trustworthiness that are listed can be grouped in several ways. One grouping might be:

- performance characteristics (accuracy),
- deployment characteristics (reliability, robustness, safety, resilience),
- adversarial characteristics (security, privacy, harmful outcomes from misuse of the AI), and
- usability characteristics (explainability, interpretability, mitigation of harmful bias).

Consider moving to a higher abstraction for the Framework to elicit the trust characteristics across a range of contexts; for example, for an Object Detection AI, mean Average Precision (mAP) is usually used instead of accuracy.

It may also be useful to consider if the organization currently has the capacity to create a trustworthy AI system, and if not, what are the steps necessary to build organizational capacity to do so? There is a large gap between creating an AI system that performs a task under ideal conditions and one that can

do so in the context of the real world. Specifically, for the real world – if deployment and acceptance is to take place, system developers (and the framework that guides them) will need to go beyond technical aspects and include a broader focus on socio-technical aspects to achieve trust required for deployment.

Suggested Change:

First, we suggest using a higher level of abstraction to guide the definition of trust characteristics for the Framework. For example, there are many additional performance characteristics, beyond accuracy, e.g. mAP, precision / recall, etc., and many additional deployment characteristics, e.g. uncertainty quantification, and so on. By moving to a higher level of abstraction, one may be better able to elicit the characteristics of trustworthiness relevant for the context of a given AI system.

More broadly, consider not just the characteristics of trust in the AI system, but also the ability of the organization to build a trustworthy system. For example, does the organization have a documented risk appetite statement that enables standardization in risk decisions in adopting or building AI systems? If not, it may be unlikely that the organization can be trusted with its AI, even if the AI itself has many of the trust characteristics.

3. How organizations currently define and manage principles of AI trustworthiness and whether there are important principles which should be considered in the Framework besides: Transparency, fairness, and accountability;

Comment:

To implement principles like these, one will need to be able to measure them within the Framework, either at a quantitative level, similar to the characteristics of trustworthiness, or at a qualitative level. For example, accountability could be measured by organizations having a documented governance structure where accountability is chartered by role.

Suggested change:

NIST may consider the use of a "Maturity Model-like" set of criteria to help organizations scale and adapt to properly account for the trustworthiness of AI and its use. This will allow for consistent qualitative measurement.

4. The extent to which AI risks are incorporated into different organizations' overarching enterprise risk management—including, but not limited to, the management of risks related to cybersecurity, privacy, and safety;

Comment:

The connection to ERM is well stated. Note, however, that there are additional risks interdependencies. For example, talent recruitment and retention is another critical risk to consider given the degree of technical complexity and demands of AI.

Suggested change:

We recommend the addition of other risk interdependencies here such as talent recruitment and retention. Additional ties of interconnectivity with an ERM portfolio could include strategy (e.g. mergers and acquisition), supply chain risk management (e.g. assessing the use of AI related product liability), and ethics (e.g ethical implementation of the technology).

Furthermore, we suggest including guidance on what is similar and what needs to be different based on domain of application. For example, ERM in the context of AI for electricity grid anomaly detection is very different from risks for AI for criminal justice sentencing.

5. Standards, frameworks, models, methodologies, tools, guidelines and best practices, and principles to identify, assess, prioritize, mitigate, or communicate AI risk and whether any currently meet the minimum attributes described above;

Comment:

A brief set of pointers to prior work follows.

1. Explain and then speculate about the overall system
    a. What problem are we solving and for whom? Is AI the right solution for the problem? Why?
    b. Use a set of ethics to support the team in this work such as the
        i. DoD's Principles for Ethical AI DOD Adopts Ethical Principles for Artificial Intelligence > U.S. Department of Defense > Release
        ii. Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities | U.S. GAO
        iii. Awesome AI Guidelines on GitHub from EthicalML: https://github.com/EthicalML/awesome-artificial-intelligence-guidelines
    c. Consider what is interesting about this system to potential adversaries?
    d. Consider what access adversaries might gain? What systems are connected?
    e. Conduct speculative activities:
        i. Checklist to prompt intentional, uncomfortable conversations: Designing Ethical AI Experiences (Carnegie Mellon University, Software Engineering Institute): https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=636620
        ii. Harms modeling (Microsoft): https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/harms-modeling/
        iii. Abusability Testing: UX in the Age of Abusability. The role of Composition, Collaboration, and Craft in building ethical products. By Dan Brown. Sep 18, 2018. https://greenonions.com/ux-in-the-age-of-abusability-797cd01f6b13
        iv. Abusability Testing: Article describing the Abusability Testing workshop organized by Anna Abovyan, Theora Kvitka and Allison Cosby of the Pittsburgh IxDA Chapter for World Interaction Design Day 2019. http://blog.daed.com/?p=2835
        v. "Black Mirror" Episodes: Black Mirror, Light Mirror: Teaching Technology Ethics Through Speculation. By Casey Fiesler. Oct 15,

2018. https://howwegettonext.com/the-black-mirror-writers-room-teaching-technology-ethics-through-speculation-f1a9e2deccf4

vi. Activity for individuals to identify their own bias - Implicit Association Test (IAT) (Harvard University): https://implicit.harvard.edu/implicit/takeatest.html

2. Data inspection
    a. Deeply inspect data and combinations of data that may result in unintended negative consequences and opportunities for misuse and abuse.
    b. Datasheets for Datasets - Paper on arxiv by Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, Kate Crawford: https://arxiv.org/abs/1803.09010
        i. Template for Datasheets for Datasets (Microsoft): https://www.microsoft.com/en-us/research/publication/datasheets-for-datasets/
        ii. Markdown for Datasheets for Datasets by JRMeyer (GitHub — includes LaTeX version): https://github.com/JRMeyer/markdown-datasheet-for-datasets
    c. Ethics checklist for data scientists by Deon: https://deon.drivendata.org/
    d. Responsible AI Toolkits (Microsoft): https://www.microsoft.com/en-us/ai/responsible-ai-resources
    e. Data Nutrition Project (Empowering data scientists and policymakers with practical tools to improve AI outcomes): https://datanutrition.org/
3. Algorithms, Models and Training Methods
    a. Deeply inspect algorithm choices, model development and training methods and speculate about unintended negative consequences and opportunities for misuse and abuse.
    b. Model Cards
        i. Model Cards for Model Reporting paper on ACM by Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. - https://dl.acm.org/doi/abs/10.1145/3287560.3287596
        ii. Templates for Model Cards (Google): https://modelcards.withgoogle.com/model-reports

6. How current regulatory or regulatory reporting requirements (e.g., local, state, national, international) relate to the use of AI standards, frameworks, models, methodologies, tools, guidelines and best practices, and principles;

No comment.

7. AI risk management standards, frameworks, models, methodologies, tools, guidelines and best practices, principles, and practices which NIST

should consider to ensure that the AI RMF aligns with and supports other efforts;

Comment:

We concur with the ERM-based approach for risk management here, as it recognizes the interdependency of AI related risks with others in the ERM risk portfolio.

Suggested change:

An additional document to assist NIST and its readers in the development of ERM policies and practices could include SEI's OCTAVE FORTE. https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=644636

8. How organizations take into account benefits and issues related to inclusiveness in AI design, development, use and evaluation—and how AI design and development may be carried out in a way that reduces or manages the risk of potential negative impact on individuals, groups, and society.

Comment:

AI systems learn from examples, so it helps to have a diverse team that can bring different lenses to a problem and identify appropriate datasets to train the AI system on. It naturally follows that assembling a team with different backgrounds that can speak to different aspects of the problem will result in a better selection of datasets. AI teams need to be informed by a range of cultures, experiences, and how team members think about the world and the heuristics they use to solve problems. A team can be made up of members with diverse backgrounds, but if all the team members are engineers, they will approach the problem space in the same way. Teams need to explore what it would mean to partner with a policy maker or a philosopher and how those unique perspectives would drive solutions that would be ethical and implementable.

9. The appropriateness of the attributes NIST has developed for the AI Risk Management Framework. (See above, ‘‘AI RMF Development and Attributes’’);

No comment.

10. Effective ways to structure the Framework to achieve the desired goals, including, but not limited to, integrating AI risk management processes with organizational processes for developing products and services for better outcomes in terms of trustworthiness and management of AI risks. Respondents are asked to identify any current models which would be effective. These could include—but are not limited to—the NIST Cybersecurity Framework or Privacy Framework, which focus on outcomes, functions, categories and subcategories and also offer options

for developing profiles reflecting current and desired approaches as well as tiers to describe degree of framework implementation; and

Comment:

Positioning the framework as a continuous learning process (see, for example Kolb's experiential learning model) can help to introduce the notion that everyone has a role to play in learning about the evolution of AI systems, the risks that emerge, and strategies for addressing them. By focusing the framework on learning toward the desired systems outcomes (i.e., systems that are trustworthy, secure, resilient, etc.) it broadens the aperture to include multiple approaches for how to reach end states, rather than focus on a single approach adopted by individuals with fixed roles and skills. Additionally, change management will be a critical part of adopting new risk management approaches as AI systems have several inherent differences from traditional software risk management and thus, Kotter's change management model might also prove useful.

11. How the Framework could be developed to advance the recruitment, hiring, development, and retention of a knowledgeable and skilled workforce necessary to perform AI-related functions within organizations.

Comment:

An assumption often exists that someone – a machine learning researcher, the CEO of an industry company, an expert – knows exactly how to manage AI-related risks in all contexts, but they don't work at the organization who needs the answers. The truth is that today, much about the implementation of AI systems is still in the artisan phase - including risk management. Applying new algorithms to real-world problems and real-world datasets is hard and it's challenging to know the risks that will emerge over time.

More: https://insights.sei.cmu.edu/blog/5-ways-to-start-growing-an-ai-ready-workforce/

Suggested Change:

Structuring the framework to foster curiosity and acknowledge the inherent complexity in risk management processes can help to encourage organizations to think broadly about recruitment and workforce diversity. A single person cannot cover all potential risks, and instead organizations should focus on identifying individuals who can reach across different boundaries within a system to track down an answer.

12. The extent to which the Framework should include governance issues, including but not limited to make up of design and development teams, monitoring and evaluation, and grievance and redress.

Comment:

Donella Meadows, key systems thinking leader said, "Pay attention to what is important, not just what is quantifiable." Governance structures and issues for AI systems must take into account what is important - and certainly the people that create and develop systems as well as system evaluators are critical to the integrity and responsibility of systems. Such teams play a role in mitigating potential risks, challenging assumptions, and are themselves - a likely system vulnerability. The framework should acknowledge governance and guide how continuous governance structures should both be constructed and supported over time.

Contact Us

Software Engineering Institute
4500 Fifth Avenue, Pittsburgh, PA 15213-2612

Phone: 412/268.5800 | 888.201.4479
Web: www.sei.cmu.edu
Email: info@sei.cmu.edu

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

This report was prepared for the SEI Administrative Agent AFLCMC/AZS 5 Eglin Street Hanscom AFB, MA 01731-2100

DM21-0803