Computing Research Association - Industry (CRA-I) and the Computing Community Consortium (CCC) Response to the Second Draft of the National Institute of Standards and Technology (NIST) Artificial Intelligence Risk Management Framework

September 2022

David Danks (University of California, San Diego) and Kathy Meier-Hellstern (Google)

Response to Request for Information (RFI) on the Second Draft of the NIST Artificial Intelligence Risk Management Framework (AI RMF or Framework): https://www.nist.gov/system/files/documents/2022/08/18/AI_RMF_2nd_draft.pdf

The following is a joint response from the Computing Research Association (CRA)'s CRA-Industry (CRA-I) and the Computing Community Consortium (CCC). Overall, version 2.0 of the NIST RMF is quite helpful and describes some valuable approaches. There are, however, some inconsistencies in terminologies that make the document confusing. In some cases, the authors introduce slightly new terminology, where there are already three or alternatives available. And even within the document, several frameworks are presented that have partial overlap, and that give the appearance of being forced together. These multiple sets of terminology make it difficult to take away a single core message. The Playbook also mixes together (a) desirable end-states; and (b) desirable processes. It would be helpful if it was clear about which recommendations are goals vs. ways to reach those goals.

Please see our specific comments below.

- Section 1.1 - A useful mathematical representation of the data interactions that drive the AI system's behavior is not fully known, which makes current methods for measuring risks and navigating the risk-benefits tradeoff inadequate. AI risks may arise from the data used to train the AI system, the AI system itself, the use of the AI system, or interaction of people with the AI system.
    - This is misleading because it implies the only way to measure risks is by understanding the inner workings of the models. It ignores other well-established approaches to risk management such as input/output adversarial testing.
- There are several places that note that the word "risk" should be understood to include both harms and benefits. However, almost all of the substantive elements of the RMF and Playbook are harm-centric. Given the importance of considering benefits, as well as the balancing act required between benefits and harms, there should be a systematic review of the RMF and especially the Playbook to ensure that uses of the word 'risk' are actually neutral between benefits and harms.

- Section 1.1 - While views about what makes an AI technology trustworthy differ, there are certain key characteristics of trustworthy systems. Trustworthy AI is valid and reliable, safe, fair and bias is managed, secure and resilient, accountable and transparent, explainable and interpretable, and privacy-enhanced.
    - The section is titled “Trustworthy and Responsible AI”, but you don’t define Responsible AI. Are you using the terms interchangeably?
    - What is “fair and bias”?
    - Terms are not consistent with other organizations. We would think this document would want to align with a set of AI principles, for example from OECD.[1] Why create another list? (And this list seems to be more scattered)
    - This version of the RMF continues to largely equate “trustworthy” with "risk minimizing." These two features are not the same, however, as trustworthiness requires much more. (Moreover, note that the term ‘risk minimization’ also suggests that the focus should be harm reduction, rather than increasing benefits.) Unfortunately, the language of the current version of the RMF moves quickly between these two terms. Given the importance of trustworthy AI efforts, both inside and outside of the US Government, it is critical that the RMF *not* equate these two terms, even implicitly. Efforts to develop trustworthy AI will need to go beyond the RMF and Playbook as currently constituted, and these documents should be explicit about that additional work.
- Section 2, Figure 1
    - It is critical to perform “testing” and analysis of collected training data to understand what types of biases or harms are in the data. The outcome of these studies should be documented in transparency artifacts like Data Cards.[2] An important aspect of creating a Data Card is documentation of the data analysis and curation that was performed on the training data. Quantifying where data is sourced from, sensitive material contained in the data, and how it relates to social groups represented in data is a needed step for anticipating, mitigating, and documenting representational harms and other downstream Trust/Responsibility concerns.
    - Key attributes of the AI model should be documented using Model Cards.[3] Model Cards are short documents accompanying trained machine learning models that provide benchmarked evaluation in a variety of conditions and disclose the model architecture, the context in which models are intended to be used, details of the performance evaluation procedures, and other relevant information. Model Cards can be completed after the model is built and evaluations are made.

[1] https://www.oecd-ilibrary.org/science-and-technology/oecd-framework-for-the-classification-of-ai-systems_cb6d9eca-en;jsessionid=yBR_x-d2ykowqE4dJtmlF9CJPLLzEn2HW4gV7RgN.ip-10-240-5-68

[2] https://arxiv.org/abs/2204.01075

[3] https://arxiv.org/pdf/1810.03993.pdf

    - Unsurprisingly (given NIST's mission), the RMF focuses primarily on TEVV, but that omits some key parts of developing in risk minimization/managing ways. For example, they include "Plan & design" in Figure 1, but the TEVV operationalization is "audit & impact assessment" (which are focused on post-deployment activities, rather than appropriate measures during pre-deployment design phase).
    - Section 3: Framing Risk - AI risk management is about offering a path to minimize potential negative impacts of AI systems, such as threats to civil liberties and rights, as well as pointing to opportunities to maximize positive impacts.
    - We don't think of risk management as pointing to opportunities. And there is no additional mention of opportunities in the rest of the document. If this document wants to talk about opportunities, that would be better placed in a separate discussion.
- Figure 4 - The terms are not aligned with other standards. It would make more sense to follow one of these. The use of (yet another) set of terms is confusing, especially because all of them are conveying similar concepts
    - Table 1 - Following from the comments in Figure 4, the document introduces yet another set of terms. (Why?). And we don't think some of the equivalencies are correct.
    - Table 1 (mapping NIST RMF to OECD, EU AI Act, and EO 13960 elements) is really helpful. Unfortunately, we are not sure that they did it correctly. In particular, "Privacy-enhanced" doesn't really line up with the OECD emphasis on "Human values"
- Section 4.1 - Deployment of AI systems which are inaccurate, unreliable, or non-generalizable to data beyond their training data (i.e., not robust) creates and increases AI risks and reduces trustworthiness.
    - We think of robustness in terms of small perturbations in the input lead to small perturbations in the output…. (or in a case of fairness, if we change a parameter that "should not matter", e.g., change the doctor, he to the doctor, she, the model accuracy should not be affected.)
- Section 4.1 - Reliability – "ability of an item to perform as required, without failure, for a given time interval, under given conditions"[4] is a goal for overall correctness of model operation under the conditions of expected use and over a given period of time, to include the entire lifetime of the system.
    - What is the definition of failure in this context? Is the document talking about accuracy or precision?
- Section 4.3 - "Fair – and Bias Is Managed"

[4] https://www.iso.org/obp/ui/#iso:std:iso-iec:ts:5723:ed-1:v1:en

  - What is meant here? We have not seen this phrasing. Are the authors trying to convey something special? There are standard terms for many types of fairness - why are we not using those?
- Section 4.4 - Secure and Resilient
  - We feel like the definition of resilience needs to be expanded upon with examples.
- Section 5 - Organizations and other users of the Framework are encouraged to periodically evaluate whether the AI RMF has improved their ability to manage AI risks, including but not limited to their policies, processes, practices, implementation plans, indicators, and expected outcomes.
  - This statement seems to be crying out for some type of metric or quantification. We would love to see some thoughts on this in the draft.
- Figure 5 seems to have some overlap with Figure 1. It reads like a second framework has been "tacked on", and is yet another layer of jargon.
- Table 2 looks like it has overlap with some concepts in Figure 2.