Socure Feedback on NIST Artificial Intelligence Risk Management Framework Draft #1

April 28, 2022

Introduction

Socure applauds the government's commitment to provide guidance for Artificial Intelligence (AI) that promotes inclusive and accessible, open and transparent, consensus-based, globally relevant, and non-discriminatory principles. Socure additionally agrees that the use of AI has the ability to uplift and empower people and to lead to new services, support, and efficiencies for people and society.

It is imperative that, as the government continues to refine its approach to AI, it establishes the right high-level characteristics of an AI system from which other risk analysis and management is drawn. This overarching consideration leads to the below comments intended to strengthen the conceptual framework from which the AI Risk Management Framework (RMF) will be drafted.

Overall

- **Issue**: Metrics are difficult to understand and implement, and measuring the wrong thing or in the wrong way can undermine the best intentions.
    - **Recommendation**: As the world's premier measurement science institute, NIST is in the best position to influence the community toward appropriate use of metrics. To that end, NIST should include a metrics catalog with the AI RMF. It would provide great benefit to a community that might otherwise struggle to wade through the large number of methods to measure AI.

Definitions and Scoping

- **Issue:** The document begins by describing the changes in AI, its increased use, and reasons for unintended outcomes in the use of AI. The callout box on page 2 provides a useful definition but this is insufficient to capture the important nuances associated with the term AI. This makes it difficult to understand what is in scope and what is not. Similarly, The scope does not address the types of systems that are in or out, referring generically to AI and the provided definition.

- **Recommendation**: NIST should define AI and its related terms (e.g. machine learning to symbolic AI) and their use cases (e.g. recommendation engines, text generation, surveillance, public/private sector uses), and how each is considered in this document.

  NIST should discuss the different use types of AI and how the AI RMF applies to each. This includes narrow vs general intelligence and supervised vs unsupervised learning.

  NIST should discuss different audiences for the system itself and how the AI RMF applies to them. This includes end-user-, expert-, and researcher-facing. That is, how should risk be considered differently when an AI system's primary audience differs.

Framework Core

- **Issue**: It can be difficult to understand how different types of AI apply to different contexts.
  - Recommendation: For each core function, NIST should ensure that each category and subcategory addresses the different types and audiences of AI.

    As you evaluate each core function and categories, consider the different types of AI. Some categories and subcategories may apply to some types of AI but, for instance, not supervised ML.

    This may be best addressed through a new subcategory or clarified subcategory in the Map.2 category on page 16. If this concept was intended to be covered by "operational context," the language could be clearer.

- **Issue**: There is a great deal of overlap between the activities of the core functions. When trying to apply categories and subcategories to the functions, overlap becomes more significant. Some might find Govern and Manage to be nearly synonymous. Managing involves mapping and measuring.
  - **Recommendation**: More clearly define how these functions relate to each other. Creating a better graphic showing the process aspects of the functions would help, perhaps with something like Map (defines) Measure (enables) Manage (executes), all of which is overlaid with Governance.

    Ensure that categories and subcategories do not stray into other functions. For example, Map.4 on page 16 includes a subcategory that seems to include no Map-related activity but activities that encompass all of Measure, Manage, and Govern:
    - Measure: "Benefits of the AI system outweigh the risks,"
    - Map: "and risks can be assessed and managed."
    - Govern: "Ideally, this evaluation should be conducted by an independent third party or by experts who did not serve as front-line developers for the

system, and who consults experts, stakeholders, and impacted communities."

- **Issue**: Consistent with the prior comment, it's particularly difficult to distinguish between the categories and subcategories in Map and Govern. For example, in the introduction to the Govern function, it's quite confusing to say "Governance is designed to ensure risks and potential impacts are identified, measured, and managed effectively and consistently." This makes it sound as though the Govern function is about doing those things (identifying, measuring, and managing risks), while it's actually about ensuring the Map, Measure, and Manage functions have an organizational structure that supports them.
    - **Recommendation**: Be more explicit that Manage is about making the decisions while Govern is about organizational structure to support the other functions. To be clear, this is a recommendation about careful wording and thorough descriptions and not restructuring the framework.

- **Issue**: Some subcategories are very wordy (Gov.5.1), cover multiple domains of activity (Gov.3.1), and have inconsistent format (Manage.1.1).

  In the AI RMF, the Measure function includes a subcategory stating "Accuracy, reliability, robustness, resilience (or ML security), explainability and interpretability, privacy, safety, bias, and other system performance or assurance criteria are measured, qualitatively or quantitatively."

  Compare this to the more focused subcategories in the NIST CSF: "External information systems are cataloged," "Data-at-rest is protected," and "Notifications from detection systems are investigated."

    - **Recommendation**: Either simplify subcategories by tightening language or splitting them into multiple, more streamlined subcategories

      The longest subcategory in the NIST CSF is RS.AN-5: "Processes are established to receive, analyze and respond to vulnerabilities disclosed to the organization from internal and external sources (e.g. internal testing, security bulletins, or security researchers)

      Compare this to the longest in the AI RMF, in Gov.3: "Decision making throughout the AI lifecycle is informed by a demographically and disciplinarily diverse team, including internal and external personnel. Specifically, teams that are directly engaged with identifying design considerations and risks include a diversity of experience, expertise and backgrounds to ensure AI systems meet requirements beyond a narrow subset of users."

      A better approach might be 4 subcategories:

i. Internal decision making teams are diverse in demographics
ii. Internal decision making teams are diverse in expertise and background
iii. External advisory teams are diverse in demographics
iv. External advisory teams are diverse in expertise and background

- **Issue**: A lack of informative references or options for finding more prescriptive guidance or recommendations.
    - **Recommendation**: NIST should include informative references for each subcategory, consistent with the approach in the CSF. This could help establish an understanding of quality frameworks for addressing each subcategory's needs. Understanding control sets are not as well developed in AI as they are in cybersecurity, there is still value in pointing to resources. These could also help ground the subcategories: if there are no good references, it may not be a subcategory ready for inclusion. If there are too many, it may be too large a subcategory.
    - **Recommendation**: Consider using the MRM regulatory framework in the financial services industry as an informative reference as it is in use currently and works well with concrete standards and quantifiable results for statistical models. The general steps and structure are a good model, as is focus on quantitative metrics, even if the specifics of the MRM itself aren't what would go in the AI RMF.

- **Issue**: NIST is normative in some subcategories.
    - **Recommendation**: To the extent possible, avoid making normative statements (e.g., "ideally"). If it is an objective approach to risk management, include it. Otherwise, stick with the more objective–if not grammatically pleasant!–"passive voice" approach of the CSF.

- **Issue**: Some language in subcategories is insufficiently concrete for risk management purposes
    - **Recommendation**: Use verbs consistent with management processes, even if it means repeating a small set of them.

      For example, Map.4.2 uses "elucidated." Elucidation is quite difficult to build a control structure around.

- **Issue**: Insufficient balance between benefits and costs, focusing instead on risks and harms.
    - **Recommendation**: Adjust some categories and subcategories to drive a consideration of value rather than just risk or harm. This tracks with multiple workshop comments about how we don't have sufficient data on benefits.

Adjust Map.4 from “Risks and harms” to “Benefits and costs to individual…”, reword subcategories as appropriate, and add a new subcategory to address ensuring benefit to users outweighs potential costs to users

Create a category in Measure about understanding the created value overall to various stakeholders and subcategories regarding different types of stakeholders.

Adjust Manage.1 to be about a fair assessment of net value and not just harm. Adjust Manage.1.1 to address this and reconsider whether Manage.1.2 should include prioritization based on risk (as stated) and also opportunity cost, i.e., what benefits do we lose if we can’t find a way to effectively mitigate high priority risks.

Map

- **Issue**: Map.1.3 states that “The organization’s mission and relevant goals for the AI technology are understood.” This is useful, but doesn’t capture the broader set of core principles that often drive decision making in organizations.
    - **Recommendation**: Include in this subcategory about aligning to organizational principles and other risk management activities, like privacy.
- **Issue**: Map.4 largely focuses on the risks of “overdoing” AI. There is a set of risks associated with “underdoing” AI that should be managed. Specifically, a lack of mechanisms to ensure sufficient coverage of populations can result in poor results for some groups, and often predictably underserved groups.
    - **Recommendation**: Create a subcategory under Map.4: “Ensure systems are developed with redundancy and data sourcing mechanisms that ensure wide population coverage to expand access to potential users, especially those that may be systemically underrepresented in common data sources.”.

Measure

- **Issue**: Measure.2 is overloaded in the first subcategory. There are so many characteristics, it risks making measurement of any given one intractable.
    - **Recommendation**: Calling out the characteristics is helpful, as is avoiding specific methods for measuring, but separate the characteristics into distinct subcategories, at least one for each characteristic subgroupings in Section 5.
- **Issue**: Measure.1 captures the idea that you have to measure how you measure. But it does not capture the idea of representative groups, and this is a worthwhile distinct call out.
    - **Recommendation**: Consider a subcategory under Measure.1: “When developing evaluation techniques, impacted individuals are represented appropriately.”
- **Issue**: The Measure function overlooks the importance of measuring inputs to systems, which is a prerequisite to successful performance of AI systems.

    - **Recommendation**: Create a subcategory, perhaps in Measure.2: “Data sources are evaluated for freshness, accuracy, bias, and latency.”
-

Manage

- **Issue**: The function name manage is a bit confusing as the document is about risk management
    - **Recommendation**: Consider changing to “Control” or “Execute” or the like.

Govern

- **Issue**: Organizational principles should be considered in governing AI systems.
    - **Recommendation**: Create a new subcategory in Govern about AI oversight to match organizational principles. This likely fits best in Govern.1.
- **Issue**: Both subcategories in Gov.6 are double barreled.
    - **Recommendation**: Split out 3rd party data and 3rd party AI from each
- **Issue**: Some categories are related and could be strengthened by combining them.
    - **Recommendation**:
        - Combine Gov.1 and Gov.2
        - Combine Gov.3 and Gov.5
- **Issue**: The Govern function does not capture the importance of an organizational structure that separates development and evaluation.
    - **Recommendation**: Create a new subcategory, perhaps under Govern.2: “Development and evaluation are sufficiently separated in the organizational structure to allow independence of goal setting and influence over management.”
- **Issue**: The Govern function addresses DE&I, accessibility, and cultural considerations, but does not specifically address these in the evaluation process, rather using more vague language.
    - **Recommendation**: Create a new subcategory, perhaps under Govern.5: “Ensure appropriate diversity as part of the evaluation process.”
- **Issue**: Govern.6 does not sufficiently address the various roles of third parties, which is necessary to provide proper governance over each.
    - **Recommendation**: Adjust the first subcategory under govern.6 to distinguish between 1) onboarding, 2) ongoing monitoring of performance as a 3rd party organization (financial solvency, reputational, etc.), and 3) ongoing monitoring of the data itself for consistency and accuracy.

Practical Guide

- **Issue**: A lack of practical guidance to make the AI RMF more understandable and implementable by the community.

- **Recommendation**: Produce the practical guide. The practical guide should discuss the tradeoffs associated with different governance schemes, like model cards vs model profiles and evaluation approaches.

Other Considerations: Privacy

- **Issue**: Privacy is a critical consideration that is given limited real estate in the current AI RMF draft. While there already is the NIST Privacy Framework, the AI RMF would benefit from additional privacy language and more concrete hooks to the Privacy Framework. Some additional suggestions:
    - **Recommendation**: Rather than focusing on data minimization, focus on contextual understanding of nuanced, risk-based controls.
        - For example, sometimes we need data on protected classes in order to conduct bias testing. Minimizing away data may achieve one privacy goal only to undermine broader goals. This may not fit cleanly in the Framework Core, but could be part of the earlier narrative.
        - Include privacy as a stakeholder as part of ML governance. Include explicitly in Map.1 and perhaps elsewhere.