AI Risk Management Framework Concept Paper

13 December 2021

NOTE TO REVIEWERS

This Artificial Intelligence Risk Management Framework (AI RMF) concept paper incorporates input from the Notice of Request for Information (RFI) released by the National Institute of Standards and Technology (NIST) on July 29, 2021[1], and discussions during the workshop, "Kicking off NIST AI Risk Management Framework," held October 19-21, 2021.[2]

Feedback on this paper will inform further development of this approach and the first draft of the AI RMF for public comment. NIST intends to publish that initial draft for public comment in early 2022, as well as to hold additional workshops, with the goal of releasing the AI RMF 1.0 in early 2023.

NIST would appreciate feedback on whether the concept proposed here is a constructive approach for the AI RMF. While additional details will be provided in the more extensive discussion drafts of the framework, NIST welcomes suggestions now about this approach as well as details and specific topics reviewers would like to see included in the upcoming first draft of the AI RMF. Specifically, NIST requests input on the following questions:

- Is the approach described in this concept paper generally on the right track for the eventual AI RMF?
- Are the scope and audience (users) of the AI RMF described appropriately?
- Are AI risks framed appropriately?
- Will the structure – consisting of Core (with functions, categories, and subcategories), Profiles, and Tiers – enable users to appropriately manage AI risks?
- Will the proposed functions enable users to appropriately manage AI risks?
- What, if anything, is missing?

Please send feedback on this paper to **AIframework@nist.gov** by **January 25, 2022**.

[1] https://www.nist.gov/itl/ai-risk-management-framework/ai-rmf-development-request-information
[2] https://www.nist.gov/news-events/events/2021/10/kicking-nist-ai-risk-management-framework

AI Risk Management Framework Concept Paper

1 Overview

This concept paper describes the fundamental approach proposed for the National Institute of Standards and Technology (NIST) Artificial Intelligence Risk Management Framework (AI RMF or framework). The AI RMF is intended for voluntary use and to address risks in the design, development, use, and evaluation of AI products, services, and systems.

AI technologies extend algorithmic methods and automation into new domains and roles, including advising people and taking actions in high-stakes decisions. The goal of managing the risks associated with the use of AI technologies for such tasks as recommendation, diagnosis, pattern recognition, and automated planning and decision-making frames opportunities for developing and using AI in ways that will increase their trustworthiness and advance their usefulness, while addressing potential harms.

AI risk management follows similar processes as other disciplines. Nevertheless, managing AI risks presents unique challenges. An example is the evaluation of effects from AI systems that are characterized as being long-term, low probability, systemic, and high impact. Tackling scenarios that can represent costly outcomes or catastrophic risks to society should consider: an emphasis on managing the aggregate risks from low probability, high consequence effects of AI systems, and the need to ensure the alignment of ever more powerful advanced AI systems. This proposed AI RMF is an initial attempt to describe how AI risks differ from other domains, and to offer directions – leveraging a multi-stakeholder approach – for creating and maintaining actionable guidance that is broadly adoptable.

AI risk management entails challenges that reflect both the breadth of AI use cases and the quickly evolving nature of the field. The AI ecosystem is home to a multitude of diverse stakeholders, including developers, users, deployers, and evaluators. For organizations, identifying, assessing, prioritizing, responding to, and communicating risks across business and social relationships at scale can be a difficult task. These challenges are amplified by the heterogeneity of AI risks, whether in degree or in kind, which potentially include harmful bias, threats to safety, privacy, and consumer protection, among others.

A voluntary consensus-driven framework can help create and safeguard trust at the heart of AI-driven systems and business models and permit the flexibility for innovation, allowing the framework to develop along with the technology.

NIST's work on the framework is consistent with its broader AI efforts – as called for by the National AI Initiative Act of 2020, the National Security Commission on Artificial Intelligence recommendations, and the Plan for Federal Engagement in AI Standards and Related Tools – for NIST to collaborate with the private and public sectors to develop the AI RMF.

2 Scope and Audience

The NIST AI RMF is intended to serve as a blueprint for mapping, measuring, and managing risks related to AI systems across a wide spectrum of types, applications, and maturity. This resource will be organized and written in such a way that it can be understood and used by the greatest number of individuals and organizations, regardless of sector or level of familiarity with a specific type of technology. Ultimately, it will be offered in multiple formats, including interactive versions, to provide maximum flexibility by users.

The intended primary audiences are:

(1) people who are responsible for designing or developing AI systems;
(2) people who are responsible for using or deploying AI systems; and
(3) people who are responsible for evaluating or governing of AI systems.

A fourth audience who will serve as a key motivating factor in this guidance is

> (4) people who experience potential harm or inequities affected by areas of risk that are newly introduced or amplified by AI systems.

All stakeholders should be involved in the risk management process. Stakeholders may include the mission/task champion (leadership), program management, system engineer, AI developer, requirements representative, test and evaluation personnel, end user, and affected communities, depending on the application.

3 Framing Risk

Within the context of the AI RMF, "risk" refers to the composite measure of an event's probability of occurring and the consequences of the corresponding events. While some interpretations of consequence focus exclusively on negative impacts (what is the likelihood that something bad will happen?), NIST intends to use a broader definition that offers a more comprehensive view of potential influences, including those that are positive, resonating with the goals of developing and applying AI technologies to achieve positive outcomes. If handled appropriately, AI technologies hold great potential to uplift and empower people and to lead to new services, support, and efficiencies for people and society. Identifying and minimizing potential costs associated with AI technologies will advance those possibilities.

AI risk management is as much about offering a path to minimize anticipated negative impacts of AI systems, such as threats to civil liberties and rights, as it is about identifying opportunities to maximize positive impacts.

The AI RMF endeavors to lay the groundwork for a common understanding of roles and responsibilities across the AI lifecycle. It is agnostic as to whether duties are dispersed and governed throughout an organization or assumed by an individual without any organizational affiliation. By focusing on *measurable criteria* that indicate AI system trustworthiness in meaningful, actionable, and testable ways, this framework concept paper lays out the components of an effective AI risk management program. The framework is designed to be readily useful to and usable by those with varied roles and responsibilities throughout the AI lifecycle.

4 Attributes of AI RMF

NIST has developed a set of attributes based on public feedback from a recent request for information and workshop. NIST encourages comments as to whether the framework envisioned in this concept paper, as well as all future documents released during the initial development of the AI RMF, successfully meets such expectations. The following constitutes NIST's list of framework attributes:

1. Be consensus-driven and developed and regularly updated through an open, transparent process. All stakeholders should have the opportunity to contribute to and comment on the AI RMF development.
2. Be clear. Use plain language that is understandable by a broad audience, including senior executives, government officials, NGO leadership, and, more broadly, those who are not AI professionals, while still of sufficient technical depth to be useful to practitioners. The AI RMF should allow for communication of AI risks across an organization, with customers, and the public at large.
3. Provide common language and understanding to manage AI risks. The AI RMF should provide taxonomy, terminology, definitions, metrics, and characterizations for aspects of AI risk that are common and relevant across sectors.
4. Be easily usable. Enable organizations to manage AI risk through desired actions and outcomes. Be readily adaptable as part of an organization's broader risk management strategy and processes.
5. Be appropriate for both technology agnostic (horizontal) as well as context-specific (vertical) use cases to be useful to a wide range of perspectives, sectors, and technology domains.

6. Be risk-based, outcome-focused, cost-effective, voluntary, and non-prescriptive. It should provide a catalog of outcomes and approaches to be used voluntarily, rather than a set of one-size-fits-all requirements.
7. Be consistent or aligned with other approaches to managing AI risks. The AI RMF should, when possible, take advantage of and foster greater awareness of existing standards, guidelines, best practices, methodologies, and tools for managing AI risks as well as illustrate the need for additional, improved resources. It should be law- and regulation-agnostic to support organizations' abilities to operate under applicable domestic and international legal or regulatory regimes.
8. Be a living document. The AI RMF should be capable of being readily updated as technology, understanding, and approaches to AI trustworthiness and uses of AI change and as stakeholders learn from implementing AI risk management generally and this framework, in particular.

5 AI RMF Structure

The proposed structure for the AI RMF is composed of three components: 1) Core, 2) Profiles, and 3) Implementation Tiers. This structure, and intended definitions for Core, Profiles and Implementation Tiers, are similar to the structure used in the NIST Framework for Improving Critical Infrastructure Cybersecurity (Cybersecurity Framework)[3] and the NIST Privacy Framework.[4]

5.1 Core

The Core provides a granular set of activities and outcomes that enable an organizational dialogue about managing AI risk. As with the NIST Cybersecurity and Privacy Frameworks, the core is not a checklist of actions to perform. Rather, it defines key outcomes identified as helpful in addressing AI risks. The Core will comprise three elements: *functions, categories,* and *subcategories*. Figure 1 illustrates the proposed structure of the Core. Table 1 provides examples to help clarify the intent of these functions.

5.1.1 Functions

Functions organize AI risk management activities at their highest level to establish the context and enumerate, assess, treat, monitor, review, and report risk. Categories are the subdivisions of a function into groups of outcomes closely tied to programmatic needs and particular activities. Subcategories further divide a category into specific outcomes of technical and/or management activities.

NIST proposes the following functions:

1. **Map:** Context is established, and risks related to the context are enumerated.
   The purpose of this function is "to find, recognize, and describe risks"[5] posed by an AI system. The baseline information gathered as part of this function informs decisions about model management, including the decisions that an AI solution is unwarranted or inappropriate versus the status quo, per a qualitative or more formal quantitative analysis of benefits, costs, and risks, and to stop development or to refrain from deployment.

   NOTE 1: *Context* refers to the domain and intended use, as well as the scope of the system, which could be associated with a timeframe, a geographical area, social environment, and cultural norms within which the expected benefits or harms exist, specific sets of users along with expectation of users, and any other system or environmental specifications.

[3] https://www.nist.gov/cyberframework
[4] https://www.nist.gov/privacy-framework
[5] See Clause 6.4.2 Risk identification at ISO 31000 2018 Risk management — Guidelines www.iso.org/obp/ui#iso:std:iso:31000:ed-2:v1:en

NOTE 2: *Enumeration of risks* refers to understanding risks posed by the AI system to individuals, groups, organizations, and society – including but not limited to risk associated with the data, model, the way in which the AI system is used, and its impact on individuals, groups, and society.

NOTE 3: This function should be performed by a team who is sufficiently diverse and multidisciplinary, representing multiple departments of the organization, and ideally includes a sufficiently diverse set of stakeholders from outside the organization.

2. **Measure:** Enumerated risks are analyzed, quantified, or tracked where possible.
   The purpose of this function is to comprehend the nature of risk or impact and its characteristics and to facilitate the management of such risk as set forth below.
   This function forms the basis for determining how the enumerated risk should be managed.

   NOTE 4: Risk analysis and quantification may include the level of risk and involve a detailed consideration of uncertainties, tradeoffs, consequences, likelihood, events, scenarios, controls, and their effectiveness. An event can have multiple causes and consequences and can affect multiple objectives.[6]

3. **Manage:** Risks are prioritized, and either avoided, mitigated, shared, transferred, or accepted based on measured severity.
   The purpose of this function is to support decisions and to select and implement options for addressing risk. Decisions should take account of the context and the actual and perceived consequences to external and internal stakeholders, as well as interactions of the proposed system with the status quo world, and potential transitional costs that may be addressed in advance of deployment or changes in status quo (including other systems, organizational structures, etc.) that may need to be made to ensure benefits are achieved and risks minimized.

   NOTE 5: This function should address the enumerated risks but also include detecting, monitoring, and tracking for risks that were not enumerated initially and process for introducing them into the updated enumerated list of risks.

4. **Govern:** Appropriate organizational measures, set of policies, processes, and operating procedures, and specification of roles and responsibilities are in place.
   The purpose of this function is to cultivate and implement a culture of risk management and to help ensure the risk responses are effectively and consistently carried out.

   NOTE 6: Governance should be part of each function *and* a function of its own, reflecting the importance of infusing governance considerations throughout risk management processes and procedures. While governance could therefore be listed as the first function, it is listed after the first three to emphasize its importance as a continual requirement for effective AI risk management over AI system lifespan. Effective risk management cannot occur where governance is robust only in the early stages of an AI system, and not as the AI system evolves or is updated over time.

5.1.2 Categories

Categories are subdivisions of a function into groups of outcomes closely tied to programmatic needs and particular activities. See Table 1 for possible examples of categories for each of the proposed functions to better explain the intent of each function. Table 1 as presented here is not intended to be exhaustive. (NIST especially invites comments on potential categories that should be included in the AI RMF.)

5.1.3 Subcategories

Subcategories further divide a Category into specific outcomes of technical and/or management activities.

[6] See Clause 6.4.3 Risk analysis at ISO 31000 2018 Risk management — Guidelines www.iso.org/obp/ui#iso:std:iso:31000:ed-2:v1:en

5.2 Profiles

Profiles enable users to prioritize AI-related activities and outcomes that best meet an organization's values, mission, or business needs and risks. Profiles can be technical and non-technical guidance for managing AI risks for context-specific use cases. They may illustrate how risk could be managed at various stages of the AI life cycle or in sector, technology, or end use applications.

5.3 Implementation Tiers

Implementation Tiers support decision-making and communication about the sufficiency of organizational processes and resources, including engineering tools and infrastructure and engineers with appropriate AI expertise, to manage AI risks deemed appropriate for the organization or situation and achieve outcomes and activities in the Profile(s).

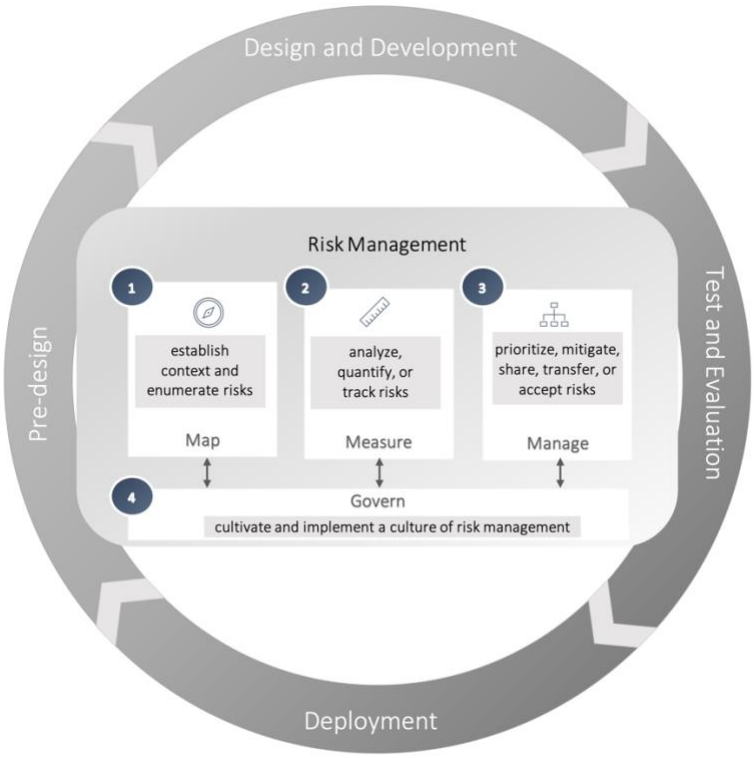

Figure 1 Risk management should be performed throughout the AI system life cycle to ensure risk management is continuous and timely. Governance is suggested as a function of its own and to be infused throughout other functions of the risk management process. Example activities for each stage of the AI life cycle follow. *Pre-design*: data collection, curation, or selection, problem formulation, and stakeholder discussions. *Design and development*: data analysis, data cleaning, model training, and requirement analysis. *Test and Evaluation*: technical validation and verification. *Deployment*: user feedback and override, post deployment monitoring, and decommissioning.

Table 1. Possible example of categories for each of the proposed functions.

ID	Category	Sub-category
Map: Context is recognized, and risks related to the context are enumerated.		
1	Context is established and understood.	
2	AI capabilities, targeted usage, goals, and expected benefits over status quo are understood.	
3	Technical, socio-technical risks[7] and harms from individual, organizational, and societal perspectives are enumerated.	
Measure: Enumerated risks are analyzed, quantified, or tracked where possible.		
	Methods and metrics for quantitative or qualitative measurement of the enumerated risks, including sensitivity, specificity, and confidence levels for specific inferences are identified and applied to the enumerated risks.	
	The likelihood of events and their consequences are assessed.	
	The effectiveness of existing security controls is evaluated.	
Manage: Enumerated risks are prioritized, mitigated, shared, transferred, or accepted based on measured severity.		
	Cost/benefit analysis (including the cost of not using AI or an assessment of whether an AI system should be developed or deployed in the first place) is performed.	
	Appropriate responses to enumerated and measured risks are identified, prepared, and implemented.	
Govern: Appropriate organizational measures, set of policies, processes, and operating procedures, and specification of roles and responsibilities are in place.		
	The resources – including engineering tools and infrastructure and engineers with appropriate AI expertise required for risk management, including contingencies – are identified.	
	Ongoing monitoring and periodic review of the risk management process and its outcomes are planned, with responsibilities clearly defined.	
	The risk management process and its outcomes are documented and reported through transparent mechanisms as appropriate.	

[7] For a discussion of AI technical and socio-technical risks see Draft-AI Risk Taxonomy www.nist.gov/system/files/documents/2021/10/15/taxonomy_AI_risks.pdf

	Decision making throughout the AI lifecycle is informed by a demographically and disciplinarily diverse team.	

6 Effectiveness of the AI RMF

Organizations are encouraged to periodically evaluate AI RMF performance against their purpose, implementation plans, indicators, and expected behavior. Sharing feedback about the AI RMF's effectiveness with NIST and others will promote continually improvement of the suitability, adequacy, and effectiveness of the AI RMF[8].

[8] See Clause 5.6 Evaluation and 5.7 Improvement at ISO 31000 2018 Risk management — Guidelines www.iso.org/obp/ui#iso:std:iso:31000:ed-2:v1:en