NIST AI Risk Management Framework

NIST AI Risk Management Framework Goals

- Be risk-based, resource-efficient, pro-innovation, and voluntary
- Be consensus-driven and developed and regularly updated through an open, transparent process
- Use clear and plain language that is understandable by a broad audience, including senior executives, government officials, non-governmental organization leadership, and those who are not AI professionals – while still of sufficient technical depth to be useful to practitioners
- Allow for communication of AI risks across an organization, between organizations, with customers, and to the public at large
- Provide common language and understanding to manage AI risks
- Be easily usable and fit well with other aspects of risk management
- Be useful to a wide range of perspectives, sectors, and technology domain
- Be outcome-focused and non-prescriptive but not one-size-fits-all requirements
- Take advantage of and foster greater awareness of existing standards, guidelines, best practices, methodologies, and tools for managing AI risks
- Be law- and regulation-agnostic
- Be a living document
- Offer a resource for improving the ability of organizations to manage AI risks to maximize benefits and to minimize AI-related harms to individuals, groups, organizations, and society

NIST AI Risk Management Framework Scope

Draft Version 2 Issued August 2022

- Address challenges unique to AI systems and encourage and equip different AI stakeholders to manage AI risks proactively and purposefully
- Describes a process for managing AI risks across a wide spectrum of types, applications, and maturity – regardless of sector, size, or level of familiarity with a specific type of technology
- Voluntary framework seeking to provide a flexible, structured, and measurable process to address AI risks prospectively and continuously throughout the AI
- Not a checklist and it is not intended to be used in isolation
- Not a compliance mechanism not intended to supersede existing regulations, laws, or other mandates
- Intended to be used by AI actors, defined by the Organisation for Economic Co-operation and Development (OECD) as "*those who play an active role in the AI system lifecycle, including organizations and individuals that deploy or operate AI*"

NIST AI Risk Management Framework
AI Actors Across the AI Lifecycle

Lifecycle	Activities	Representative Actors
Plan & design	Articulate and document the system's concept and objectives, underlying assumptions, context and requirements.	System operators, end-users, domain experts, AI designers, impact assessors, TEVV experts, product managers, compliance experts, auditors, governance experts, organizational management, end-users, affected individuals/communities, evaluators.
Collect & process data	Data collection & Processing: gather, validate, and clean data and document the metadata and characteristics of the dataset.	Data scientists, domain experts, socio-cultural analysts, human factors experts, data engineers, data providers, TEVV experts.
Build & use model	Create or select, train models or algorithms.	Modelers, model engineers, data scientists, developers, and domain experts. With consultation of socio-cultural analysts familiar with the application context, TEVV experts.
Verify & validate	Verify & validate, calibrate, and interpret model output.	
Deploy	Pilot, check compatibility with legacy systems, verify regulatory compliance, manage organizational change, and evaluate user experience.	System integrators, developers, systems/software engineers, domain experts, procurement experts, third-party suppliers with consultation of human factors experts, socio-cultural analysts, and governance experts, TEVV experts, end-users.
Operate & monitor	Operate the AI system and continuously assess its recommendations and impacts (both intended and unintended) in light of objectives and ethical considerations.	System operators, end-users, domain experts, AI designers, impact assessors, TEVV experts, product managers, compliance experts, auditors, governance experts, organizational management, end-users, affected individuals/communities, evaluators.
Use or impacted by	Use system/technology; monitor & assess impacts; seek mitigation of impacts, advocate for rights.	End-users, affected individuals/communities, general public; policy makers, standards organizations, trade associations, advocacy groups, environmental groups, civil society organizations, researchers.

NIST AI Risk Management Framework Key Definitions

- **Risk**: the composite measure of an event's probability of occurring and the magnitude (or degree) of the consequences of the corresponding events
- **Risk Management**: coordinated activities to direct and control an organization with regard to risk
- **Risk Tolerance**: The organization's or stakeholder's readiness or appetite to bear the risk in order to achieve its objectives
- **Reliability:** Ability of an item to perform as required, without failure, for a given time interval, under given conditions
- **Robustness or generalizability**: Ability of an AI system to maintain its level of performance under a variety of circumstances

NIST AI Risk Management Framework
Risk Management Challenges

- AI risks and impacts that are not well-defined or adequately understood are difficult to measure quantitatively or qualitatively
- Cannot prescribe risk tolerance – need to equip organizations to define reasonable risk tolerance, manage those risks, and document their risk management process
- Attempting to eliminate risk entirely can be counterproductive in practice – because incidents and failures cannot be eliminated – and may lead to unrealistic expectations and resource allocation that may exacerbate risk and make risk triage impractical
- Need to integrate AI risks with other critical risks

NIST AI Risk Management Framework Trustworthiness

Trustworthy AI is: valid and reliable, safe, fair and bias is managed, secure and resilient, accountable and transparent, explainable and interpretable, and privacy-enhanced

Trustworthiness Characteristics:

- Approaches which enhance AI trustworthiness can also contribute to a reduction of AI risks
- Addressing AI trustworthy characteristics individually will not assure AI system trustworthiness, and tradeoffs are always involved
- Increasing the breadth and diversity of stakeholder input throughout the AI lifecycle can enhance opportunities for identifying AI system benefits and positive impacts, and increase the likelihood that risks arising in social contexts are managed appropriately

NIST AI Risk Management Framework
Mapping of AI RMF taxonomy to AI policy documents

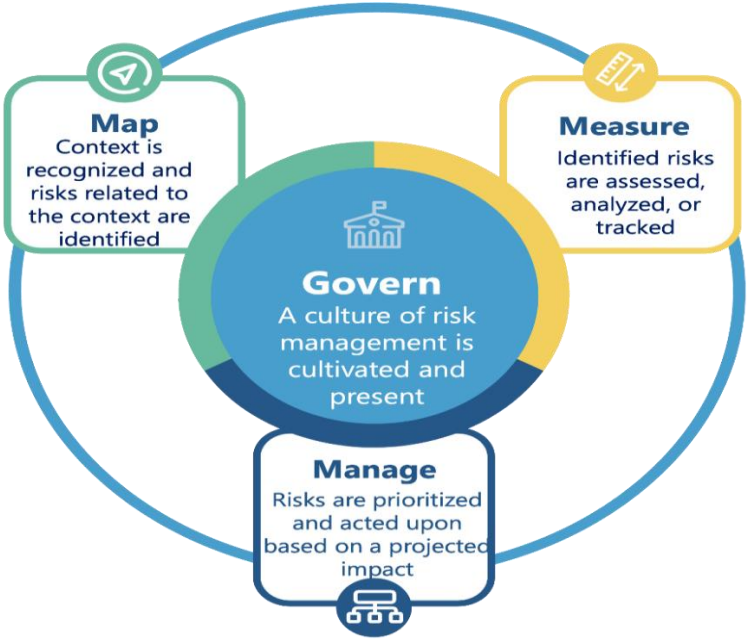
AI RMF	OECD AI Recommendation	EU AI Act (Proposed)	EO 13960
Valid and reliable	Robustness	Technical robustness	Purposeful and performance driven Accurate, reliable, and effective Regularly monitored
Safe	Safety	Safety	Safe
Fair and bias is managed	Human-centered values and fairness	Non-discrimination Diversity and fairness Data governance	Lawful and respectful of our Nation's values
Secure and resilient	Security	Security & resilience	Secure and resilient
Transparent and accountable	Transparency and responsible disclosure Accountability	Transparency Accountability Human agency and oversight	Transparent Accountable Lawful and respectful of our Nation's values Responsible and traceable Regularly monitored
Explainable and interpretable	Explainability		Understandable by subject matter experts, users, and others, as appropriate
Privacy-enhanced	Human values; Respect for human rights	Privacy Data governance	Lawful and respectful of our Nation's values

NIST AI Risk Management Framework Trusworthiness Characteristics

- **Valid and Reliable -** should consider that certain types of failures can cause greater harm – and risks should be managed to minimize the negative impact of those failures
- **Safe** - Should not, under defined conditions, cause physical or psychological harm or lead to a state in which human life, health, property, or the environment is endangered
- **Fair and Bias is Managed** - Includes concerns for equality and equity by addressing issues such as bias and discrimination
- **Secure and Resilient** - AI systems that can withstand adversarial attacks, or more generally, unexpected changes in their environment or use, or to maintain their functions and structure in the face of internal and external change, and to degrade gracefully when this is necessary
- **Transparent and Accountable** – Reflects the extent to which information is available to individuals about an AI system, if they are interacting – or even aware that they are interacting – with such a system
- **Explainable and Interpretable** - A representation of the mechanisms underlying an algorithm's operation, whereas interpretability refers to the meaning of AI systems' output in the context of its designed functional purpose
- **Privacy-Enhanced** - Norms and practices that help to safeguard human autonomy, identity, and dignity

NIST AI Risk Management Framework Core Steps

- **Govern**: Cultivate and implement a culture of risk management within organizations developing, deploying, or acquiring AI systems
- **Map**: Establish the context to frame risks related to an AI system
- **Measure**: Employ quantitative, qualitative, or mixed-method tools, techniques, and methodologies to analyze, assess, benchmark, and monitor AI risk and related impacts
- **Manage**: Entails allocating risk management resources to mapped and measured risks on a regular basis and as defined by the Govern function



NIST AI Risk Management Framework Govern Step Categories/Subcategories

GOVERN 1: Policies, processes, procedures, and practices across the organization related to the mapping, measuring, and managing of AI risks are in place, transparent, and implemented effectively.	GOVERN 1.1: Legal and regulatory requirements involving AI are understood, managed, and documented.
	GOVERN 1.2: The characteristics of trustworthy AI are integrated into organizational policies, processes, and procedures.
	GOVERN 1.3: The risk management process and its outcomes are established through transparent mechanisms and all significant risks as determined are measured.
	GOVERN 1.4: Ongoing monitoring and periodic review of the risk management process and its outcomes are planned, with organizational roles and responsibilities clearly defined.
GOVERN 2: Accountability structures are in place so that the appropriate teams and individuals are empowered, responsible, and trained for mapping, measuring, and managing AI risks.	GOVERN 2.1: Roles and responsibilities and lines of communication related to mapping, measuring, and managing AI risks are documented and are clear to individuals and teams throughout the organization.
	GOVERN 2.2: The organization's personnel and partners are provided AI risk management training to enable them to perform their duties and responsibilities consistent with related policies, procedures, and agreements.

NIST AI Risk Management Framework Govern Step Categories/Subcategories

GOVERN 3: Workforce diversity, equity, inclusion, and accessibility processes are prioritized in the mapping, measuring, and managing of AI risks throughout the lifecycle.	GOVERN 3.1: Decision-making related to mapping, measuring, and managing AI risks throughout the 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.
GOVERN 4: Organizational teams are committed to a culture that considers and communicates risk.	GOVERN 4.1: Organizational practices are in place to foster a critical thinking and safety-first mindset in the design, development, and deployment of AI systems to minimize negative impacts.
	GOVERN 4.2: Organizational teams document the risks and impacts of the technology they design, develop, or deploy and communicate about the impacts more broadly.
	GOVERN 4.3: Organizational practices are in place to enable testing, identification of incidents, and information sharing.

NIST AI Risk Management Framework Govern Step Categories/Subcategories

GOVERN 5: Processes are in place for robust stakeholder engagement.	GOVERN 5.1: Organizational policies and practices are in place to collect, consider, prioritize, and integrate external stakeholder feedback regarding the potential individual and societal impacts related to AI risks.
	GOVERN 5.2: Mechanisms are established to enable AI actors to regularly incorporate adjudicated stakeholder feedback into system design and implementation.
GOVERN 6: Policies and procedures are in place to address AI risks arising from third-party software and data and other supply chain issues.	GOVERN 6.1: Policies and procedures are in place that address risks associated with third-party entities.
	GOVERN 6.2: Contingency processes are in place to handle failures or incidents in third-party data or AI systems deemed to be high-risk.

NIST AI Risk Management Framework Map Step Categories/Subcategories

MAP 1: Context is established and understood.	MAP 1.1: Intended purpose, prospective settings in which the AI system will be deployed, the specific set or types of users along with their expectations, and impacts of system use are understood and documented. Assumptions and related limitations about AI system purpose and use are enumerated, documented, and tied to TEVV considerations and system metrics.
	MAP 1.2: Inter-disciplinary AI actors, competencies, skills, and capacities for establishing context reflect demographic diversity and broad domain and user experience expertise, and their participation is documented. Opportunities for interdisciplinary collaboration are prioritized.
	MAP 1.3: The business value or context of business use has been clearly defined or – in the case of assessing existing AI systems – re-evaluated.
	MAP 1.4: The organization's mission and relevant goals for the AI technology are understood.
	MAP 1.5: Organizational risk tolerances are determined.
	MAP 1.6: Practices and personnel for design activities enable regular engagement with stakeholders, and integrate actionable user and community feedback about unanticipated negative impacts.
	MAP 1.7: System requirements (e.g., "the system shall respect the privacy of its users") are elicited and understood from stakeholders. Design decisions take socio-technical implications into account to address AI risks.

NIST AI Risk Management Framework Map Step Categories/Subcategories

MAP 2: Classification of the AI system is performed.	MAP 2.1: The specific task, and methods used to implement the task, that the AI system will support is defined (e.g., classifiers, generative models, recommenders).
	MAP 2.2: Information is documented about the system's knowledge limits and how output will be utilized and overseen by humans.
	MAP 2.3: Scientific integrity and TEVV considerations are identified and documented, including those related to experimental design, data collection and selection (e.g., availability, representativeness, suitability), and cstruct validation.
MAP 3: AI capabilities, targeted usage, goals, and expected benefits and costs compared with the status quo are understood.	MAP 3.1: Benefits of intended system functionality and performance are examined and documented.
	MAP 3.2: Potential costs, including non-monetary costs, which result from expected or realized errors or system performance are examined and documented.
	MAP 3.3: Targeted application scope is specified, narrowed, and documented based on established context and AI system classification.

NIST AI Risk Management Framework Map Step Categories/Subcategories

MAP 4: Risks and benefits are mapped for third-party software and data.	MAP 4.1: Approaches for mapping third-party technology risks are in place and documented.
	MAP 4.2: Internal risk controls for third-party technology risks are in place and documented.
MAP 5: Impacts to individuals, groups, communities, organizations, and society are assessed.	MAP 5.1: Potential positive and negative impacts to individuals, groups, communities, organizations, and society are regularly identified and documented.
	MAP 5.2: Likelihood and magnitude of each identified impact based on expected use, past uses of AI systems in similar contexts, public incident reports, stakeholder feedback, or other data are identified and documented.
	MAP 5.3: Assessments of benefits versus impacts are based on analyses of impact, magnitude, and likelihood of risk.

NIST AI Risk Management Framework Measure Step Categories/Subcategories

MEASURE 1: Appropriate methods and metrics are identified and applied.	MEASURE 1.1: Approaches and metrics for quantitative or qualitative measurement of the most significant risks, identified by the outcome of the Map function, including context-relevant measures of trustworthiness are identified and selected for implementation. The risks or trustworthiness characteristics that will not be measured are properly documented.
	MEASURE 1.2: Appropriateness of metrics and effectiveness of existing controls is regularly assessed and updated.
	MEASURE 1.3: Internal experts who did not serve as front-line developers for the system and/or independent assessors are involved in regular assessments and updates. Domain experts, users, and external stakeholders and affected communities are consulted in support of assessments.

NIST AI Risk Management Framework Measure Step Categories/Subcategories

Category	Subcategory
MEASURE 2: Systems are evaluated for trustworthy characteristics.	MEASURE 2.1: Test sets, metrics, and details about the tools used during test, evaluation, validation, and verification (TEVV) are documented
	MEASURE 2.2: Evaluations involving human subjects comply with human subject protection requirements; and human subjects or datasets are representative of the intended population.
	MEASURE 2.3: System performance or assurance criteria are measured qualitatively or quantitatively and demonstrated for conditions similar to deployment setting(s). Measures are documented.
	MEASURE 2.4: Deployed product is demonstrated to be valid and reliable. Limitations of the generalizability beyond the conditions under which the technology was developed are documented.
	MEASURE 2.5: AI system is evaluated regularly for safety. Deployed product is demonstrated to be safe and can fail safely and gracefully if it is made to operate beyond its knowledge limits. Safety metrics implicate system reliability and robustness, real-time monitoring, and response times for AI system failures.
	MEASURE 2.6: Computational bias is evaluated regularly and results are documented.
	MEASURE 2.7: AI system resilience and security is evaluated regularly and documented.
	MEASURE 2.8: AI model is explained, validated, and documented. AI system output is interpreted within its context and to inform responsible use and governance.
	MEASURE 2.9: Privacy risk of the AI system is examined regularly and documented.
	MEASURE 2.10: Environmental impact and sustainability of model training and management activities are assessed and documented.

NIST AI Risk Management Framework Measure Step Categories/Subcategories

MEASURE 3: Mechanisms for tracking identified risks over time are in place.	MEASURE 3.1: Approaches, personnel, and documentation are in place to regularly identify and track existing and emergent risks based on factors such as intended and actual performance in deployed contexts.
	MEASURE 3.2: Risk tracking approaches are considered for settings where risks are difficult to assess using currently available measurement techniques or are not yet available.
MEASURE 4: Feedback about efficacy of measurement is gathered and assessed.	MEASURE 4.1: Measurement approaches for identifying risks are connected to deployment context(s) and informed through consultation with domain experts and other end users. Approaches are documented.
	MEASURE 4.2: Measurement results regarding system trustworthiness in deployment context(s) are informed by domain expert and other stakeholder feedback to validate whether the system is performing consistently as intended. Results are documented.
	MEASURE 4.3: Measurable performance improvements (e.g., participatory methods) based on stakeholder consultations are identified and documented.

NIST AI Risk Management Framework Manage Step Categories/Subcategories

MANAGE 1: AI risks based on impact assessments and other analytical output from the Map and Measure functions are prioritized, responded to, and managed.	MANAGE 1.1: Determination is made about whether the AI system achieves its intended purpose and stated objectives and should proceed in development or deployment.
	MANAGE 1.2: Treatment of documented risks is prioritized based on impact, likelihood, and available resources methods.
	MANAGE 1.3: Responses to the most significant risks, identified by the Map function, are developed, planned, and documented. Risk response options can include mitigating, transferring, sharing, avoiding, or accepting.
MANAGE 2: Strategies to maximize benefits and minimize negative impacts are planned, prepared, implemented, and documented, and informed by stakeholder input.	MANAGE 2.1: Resources required to manage risks are taken into account, along with viable alternative systems, approaches, or methods, and related reduction in severity of impact or likelihood of each potential action.
	MANAGE 2.2: Mechanisms are in place and applied to sustain the value of deployed AI systems.
	MANAGE 2.3: Mechanisms are in place and applied to supersede, disengage, or deactivate AI systems that demonstrate performance or outcomes inconsistent with intended use.

NIST AI Risk Management Framework Manage Step Categories/Subcategories

MANAGE 3: Risks from third-party entities are managed.	MANAGE 3.1: Risks from third-party resources are regularly monitored, and risk controls are applied and documented.
MANAGE 4: Responses to identified and measured risks are documented and monitored regularly.	MANAGE 4.1: Post-deployment system monitoring plans are implemented, including mechanisms for capturing and evaluating user and stakeholder feedback, appeal and override, decommissioning, incident response, and change management.
	MANAGE 4.2: Measurable continuous improvement activities are integrated into system updates and include regular stakeholder engagement.