AI Risk Management Framework: Second Draft

August 18, 2022

Notes for Reviewers: Call for comments and contributions

The AI Risk Management Framework

This second draft of the NIST Artificial Intelligence Risk Management Framework (AI RMF, or Framework) builds on the initial March 2022 version and the December 2021 concept paper. It reflects and incorporates the constructive feedback received.

The AI RMF is *intended for voluntary use* to address risks in the design, development, use, and evaluation of AI products, services, and systems. AI research and development, as well as the standards landscape, is evolving rapidly. For that reason, the AI RMF and its companion documents will evolve over time and reflect new knowledge, awareness, and practices. NIST intends to continue its engagement with stakeholders to keep the Framework up to date with AI trends and reflect experience based on the use of the AI RMF. Ultimately, the AI RMF will be offered in multiple formats, including online versions, to provide maximum flexibility.

Part 1 of the AI RMF draft explains the motivation for developing and using the Framework, its audience, and the framing of AI risk and trustworthiness.

Part 2 includes the AI RMF Core and a description of Profiles and their use.

NIST welcomes feedback on this draft to inform further development of the AI RMF. **Please send comments via email to AIframework@nist.gov by September 29, 2022.** All comments received by NIST will be made publicly available, so personal or sensitive information should not be included. Feedback will also be welcomed during discussions at a third AI RMF workshop on October 18-19, 2022. NIST plans to publish AI RMF 1.0 in January 2023.

NIST
NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

The AI Risk Management Framework (AI RMF) Playbook

In concert with the release of the AI RMF second draft, NIST seeks comments on a draft **companion AI RMF Playbook**. The Playbook provides actions Framework users could take to implement the AI RMF by incorporating trustworthiness considerations in the design, development, use, and evaluation of AI systems. The draft Playbook is based on this second draft of the AI RMF. It includes example actions, references, and supplementary guidance for "Govern" and "Map" – two of the four proposed functions. Draft material for the "Measure" and "Manage" functions will be released at a later date.

Like the AI RMF, the Playbook is intended for voluntary use. Organizations can utilize this information according to their needs and interests. NIST encourages feedback and contributions on this draft. The Playbook is an online resource, and is hosted temporarily on GitHub Pages.

The initial AI RMF Playbook will be published in January 2023. It is intended to be an evolving resource, and interested parties can submit feedback and suggested additions for adjudication on a rolling basis.

Individuals are encouraged to comment on:

1. Its relative usefulness as a complementary resource to the AI RMF.
2. Whether the guidance is actionable to meet each of the AI RMF functions, especially as related to organization size.
3. Suggested presentation alternatives for the forthcoming first online version to improve usability and effectiveness.

Feedback may be provided either as comments or as specific line-edit additions or modifications. NIST welcomes suggestions about including references to existing resources or new resources to help users of the AI RMF. Comments can be suggested for the Playbook at any time and will be reviewed and integrated on a semi-annual basis. NIST is requesting a first round of comments via email to AIframework@nist.gov by September 29, 2022. Comments also will be welcomed during discussions at a third AI RMF workshop on October 18-19, 2022, and beyond.

The NIST Trustworthy and Responsible AI Resource Center

The NIST Trustworthy and Responsible AI Resource Center will host the AI RMF, the Playbook, and related resources to provide guidance to implement the AI RMF as well as advance trustworthy AI more broadly. Contributions of additional guidance – which will constitute the bulk of the Resource Center content – are welcome at any time. Contributions may include AI RMF profiles, explanatory papers, document templates, approaches to measurement and evaluation, toolkits, datasets, policies, or a proposed AI RMF crosswalk with other resources – including standards and frameworks. Eventually, contributions could include AI RMF case studies, reviews of Framework adoption and effectiveness, educational materials, additional technical forms of technical guidance related to the management of trustworthy AI, and other implementation resources. The AI Resource Center is expected to include a standards hub and a metrics hub, along with a terminology knowledge base and relevant technical and policy documents.

Contributed guidance may address issues including *but not limited to:*

- how an industry or sector may utilize the Framework
- how smaller organizations can use the Framework
- how stakeholder engagement elements of the Framework can be put in place (including practices for involving external stakeholder communities for assessing risk, and advancing diversity, equity, and inclusiveness considerations within organizational teams that produce AI)
- how the AI RMF can be used for procurement or acquisition activities
- approaches for integrating teams across one or more parts of the AI lifecycle
- how the Framework can be used with other AI risk management guidance
- how the Framework can help address security concerns, including guarding against adversarial attacks on AI systems
- socio-technical approaches for evaluating AI system risks
- techniques for enhancing human oversight of AI system risks

Criteria for Inclusion of Contributions in the AI RMF Playbook and NIST Trustworthy and Responsible AI Resource Center

In order to be considered by NIST for inclusion in the Playbook or the NIST Trustworthy and Responsible AI Resource Center, a resource must be publicly available on the Internet. NIST welcomes free resources from for-profit entities. Pay-for resources from non-profit entities also meet the basic criteria for inclusion. If a resource meets these criteria, a description of the resource should be sent to AIframework@nist.gov.

NIST may include commercial entities, equipment, or materials in its guidance or in the NIST Trustworthy and Responsible AI Resource Center in order to support Framework understanding and use. Such identification does not imply recommendation or endorsement by NIST, nor that the entities, materials, or equipment are necessarily the best available for the purpose.

Update Schedule: The AI RMF will employ a two-number versioning system to track and identify major changes and key decisions that are made throughout its development. The first number will represent the generation of the AI RMF and all of its companion documents. The AI RMF generation will be incremented upon a major revision being made to the Framework. The most up-to-date version of all AI RMF documents will have the same generation identifier. Minor revisions will be tracked using ".n" after the generation number. The minor revision identifier will be incremented upon any edit to the document. It is possible that AI RMF resources will have different minor revision identifiers. At a high level, all changes will be tracked using a Version Control Table which identifies the history, including version number, date of change, and description of change.

Table of Contents

Part 1

Part 2

Appendices

AI Risk Management Framework
Part 1: Motivation

1. Overview

1.1. Trustworthy and Responsible AI

Remarkable surges in artificial intelligence (AI) capabilities have led to a wide range of innovations with the potential to benefit nearly all aspects of our society and economy – from commerce and healthcare to transportation and cybersecurity. AI technologies are often used to achieve a beneficial impact by informing, advising, or simplifying tasks.

Managing AI risk is not unlike managing risk for other types of technology. Risks to any software or information-based system apply to AI, including concerns related to cybersecurity, privacy, safety, and infrastructure. Like those areas, effects from AI systems can be characterized as long- or short-term, high- or low-probability, systemic or localized, and high- or low-impact. However, AI systems bring a set of risks that require specific consideration and approaches. AI systems can amplify, perpetuate, or exacerbate inequitable outcomes. AI systems may exhibit emergent properties or lead to unintended consequences for individuals and communities. 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.

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 AI RMF refers to an *AI system* as an engineered or machine-based system that can, for a given set of human-defined objectives, generate outputs such as predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy (Adapted from: OECD Recommendation on AI:2019; ISO/IEC 22989:2022).

AI systems are socio-technical in nature, meaning they are a product of the complex human, organizational, and technical factors involved in their design, development, and use. Many of the trustworthy AI characteristics – such as bias, fairness, interpretability, and privacy – are directly connected to societal dynamics and human behavior. AI risks – and benefits – can emerge from the interplay of technical aspects combined with socio-technical factors related to how a system is used, its interactions with other AI systems, who operates it, and the social context into which it is deployed.

Responsible use and practice of AI systems is a counterpart to AI system trustworthiness. AI systems are not inherently bad or risky, and it is often the contextual environment that determines whether or not negative impact will occur. The AI Risk Management Framework (AI RMF) can help organizations enhance their understanding of how the contexts in which the AI systems they build and deploy may interact with and affect individuals, groups, and communities. Responsible AI use and practice can:

» assist AI designers, developers, deployers, evaluators, and users to think more critically about context and potential or unexpected negative and positive impacts;
» be leveraged to design, develop, evaluate, and use AI systems with impact in mind; and
» prevent, preempt, detect, mitigate, and manage AI risks.

1.2. Purpose of the AI RMF

Cultivating trust by understanding and managing the risks of AI systems will help preserve civil liberties and rights, and enhance safety while creating opportunities for innovation and realizing the full potential of this technology. The AI RMF is intended to address challenges unique to AI systems and encourage and equip different AI stakeholders to manage AI risks proactively and purposefully. The Framework 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. Rather than repeat information in other guidance, the AI RMF aims to fill gaps specific to AI risks. Users of the AI RMF are encouraged to address non-AI specific issues via currently available guidance.

The AI RMF is a voluntary framework seeking to provide a flexible, structured, and measurable process to address AI risks prospectively and continuously throughout the AI lifecycle. It is intended to help organizations manage both enterprise and societal risks related to the design, development, deployment, evaluation, and use of AI systems through improved understanding, detection, and preemption. Using the AI RMF can assist organizations, industries, and society to understand and determine their acceptable levels of risk.

The AI RMF is not a checklist and it is not intended to be used in isolation. Organizations may find it valuable to incorporate the AI RMF into broader considerations of enterprise risk management.

The AI RMF is not a compliance mechanism. It is law- and regulation-agnostic, as AI policy discussions are live and evolving. While risk management practices should incorporate and align to applicable laws and regulations, this document is not intended to supersede existing regulations, laws, or other mandates.

The research community may find the AI RMF to be useful in evaluating various aspects of trustworthy and responsible AI and related impacts.

By applying recommendations in the AI RMF, organizations will be better equipped to govern, map, measure, and manage the risks of AI. Using the AI RMF may reduce the likelihood and degree of negative impacts and increase the benefits to individuals, groups, communities, organizations, and society.

Applying the Framework at the beginning of an AI system's lifecycle should dramatically increase the likelihood that the resulting system will be more trustworthy – and that risks to individuals, groups, communities, organizations, and society will be managed more effectively. It is incumbent on Framework users to apply the AI RMF functions to AI systems on a regular basis as context, stakeholder expectations, and knowledge will evolve over time and as their AI systems are updated or expanded.

NIST's development of the AI RMF in collaboration with the private and public sectors is directed – and consistent with its broader AI efforts called for – by the National Artificial Intelligence Initiative Act of 2020 (P.L. 116-283), the National Security Commission on Artificial Intelligence recommendations, and the Plan for Federal Engagement in Developing Technical Standards and Related Tools. Engagement with the broad AI community during this Framework's development informs AI research and development and evaluation by NIST and others.

Part 1 of this Framework establishes the context for the AI risk management process. Part 2 provides guidance on outcomes and activities to carry out that process to maximize the benefits and minimize the risks of AI. A companion resource, the AI RMF Playbook, offers sample practices to be considered in carrying out this guidance, before, during, and after AI products, services, and systems are developed and deployed.

1.3. Where to Get More Information

The Framework and supporting resources will be updated, expanded, and improved based on evolving technology, the standards landscape around the globe, and stakeholder feedback. As the AI RMF is put into use, additional lessons will be learned to inform future updates and additional resources.

The AI RMF and the Playbook will be supported by a broader NIST Trustworthy and Responsible AI Resource Center containing documents, taxonomies, toolkits, datasets, code, and other forms of technical guidance related to the development and implementation of trustworthy AI. The Resource Center will include a knowledge base of trustworthy and responsible AI terminology and how those terms are used by different stakeholders, along with documents that provide a deeper understanding of trustworthy characteristics and their inherent challenges.

The AI RMF provides high-level guidance for managing the risks of AI systems. While practical guidance published by NIST serves as an informative reference, all such guidance remains voluntary.

Attributes of the AI RMF

The AI RMF strives to:

1. Be risk-based, resource-efficient, pro-innovation, and voluntary.
2. Be consensus-driven and developed and regularly updated through an open, transparent process. All stakeholders should have the opportunity to contribute to the AI RMF's development.
3. 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. The AI RMF should allow for communication of AI risks across an organization, between organizations, with customers, and to the public at large.
4. Provide common language and understanding to manage AI risks. The AI RMF should offer taxonomy, terminology, definitions, metrics, and characterizations for AI risk.
5. Be easily usable and fit well with other aspects of risk management. Use of the Framework should be intuitive and readily adaptable as part of an organization's broader risk management strategy and processes. It should be consistent or aligned with other approaches to managing AI risks.
6. Be useful to a wide range of perspectives, sectors, and technology domains. The AI RMF should be universally applicable to any AI technology and to context-specific use cases.
7. Be outcome-focused and non-prescriptive. The Framework should provide a catalog of outcomes and approaches rather than prescribe one-size-fits-all requirements.
8. 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.
9. Be law- and regulation-agnostic. The Framework should support organizations' abilities to operate under applicable domestic and international legal or regulatory regimes.
10. Be a living document. The AI RMF should be 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.

2. Audience

Identifying and managing AI risks and impacts – both positive and negative – requires a broad set of perspectives and stakeholders. The AI RMF is 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*" [OECD (2019) Artificial Intelligence in Society | OECD iLibrary].

OECD has developed a framework for classifying AI actors and their AI lifecycle activities according to five key socio-technical dimensions, each with properties relevant for AI policy and governance, including risk management [OECD (2022) OECD Framework for the Classification of AI systems | OECD Digital Economy Papers]. For purposes of this framework, NIST has slightly modified OECD's classification. The NIST modification (shown in Figure 1) highlights the importance of test, evaluation, verification, and validation (TEVV) throughout an AI lifecycle and generalizes the operational context of an AI system.

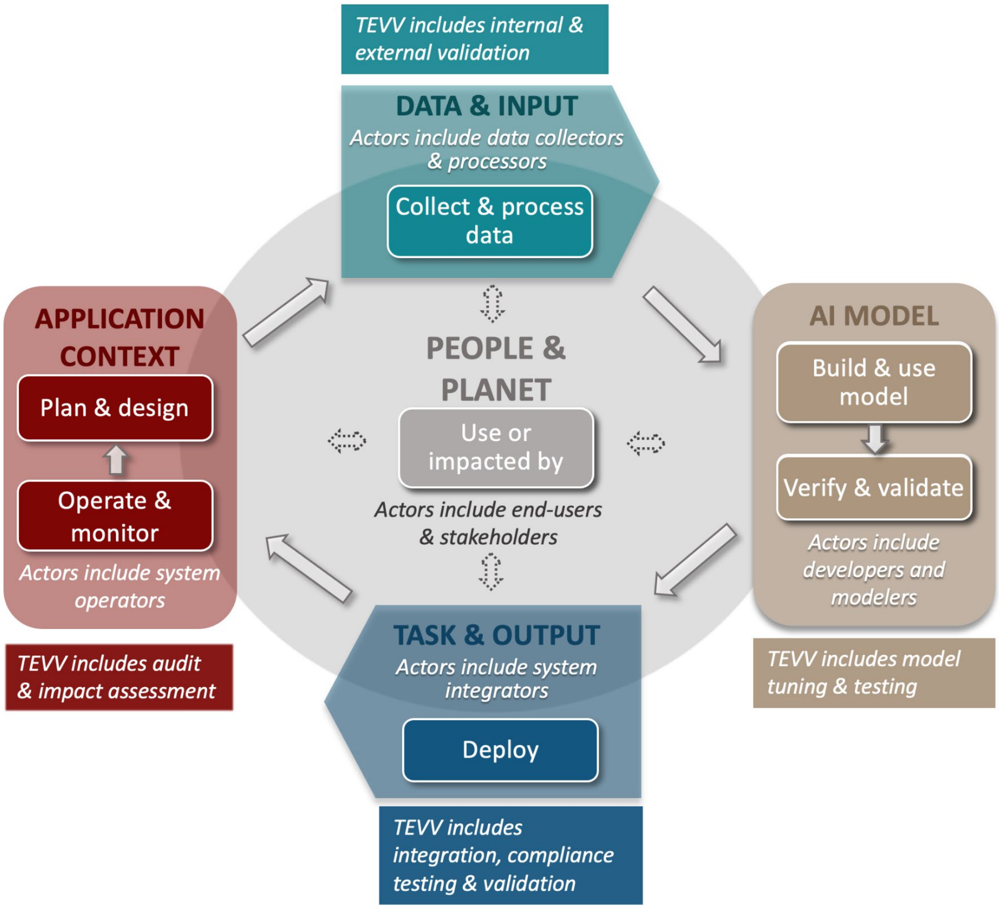
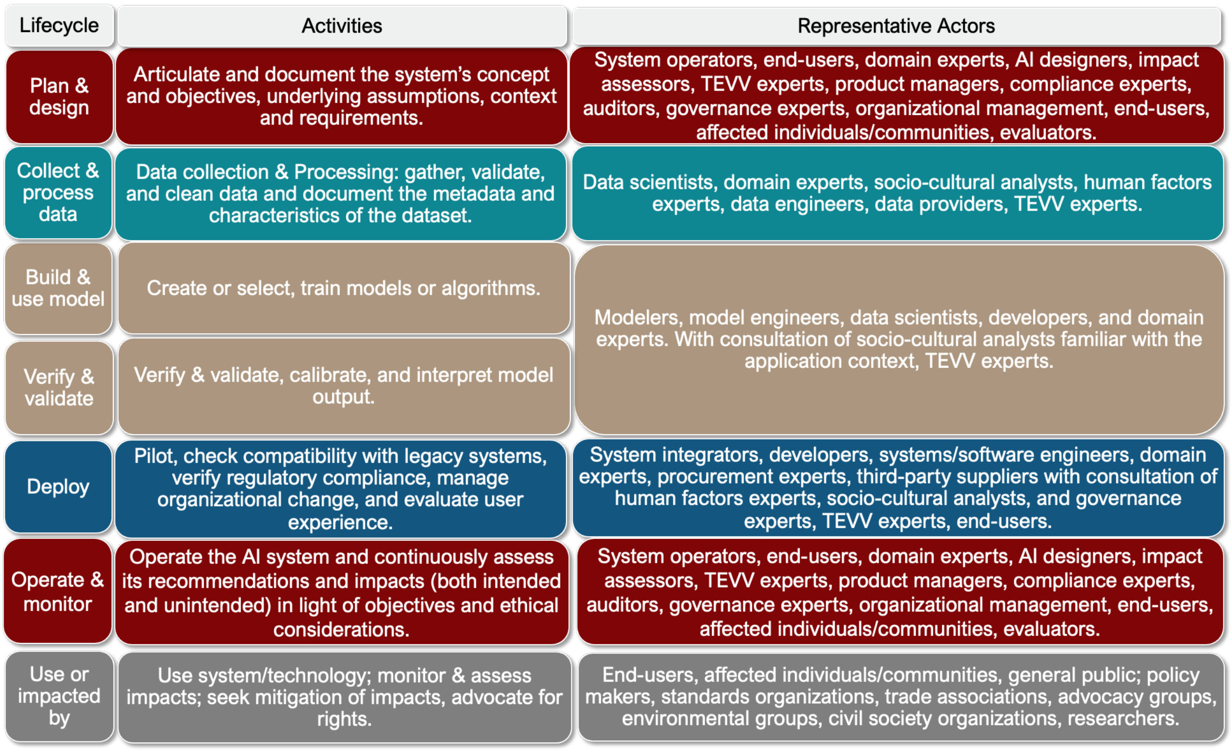

TEVV includes internal & external validation
DATA & INPUT
Actors include data collectors & processors
Collect & process data
APPLICATION CONTEXT
Plan & design
Operate & monitor
Actors include system operators
TEVV includes audit & impact assessment
PEOPLE & PLANET
Use or impacted by
Actors include end-users & stakeholders
AI MODEL
Build & use model
Verify & validate
Actors include developers and modelers
TEVV includes model tuning & testing
TASK & OUTPUT
Actors include system integrators
Deploy
TEVV includes integration, compliance testing & validation

Figure 1: Lifecycle and Key Dimensions of an AI System. Modified from OECD (2022) OECD Framework for the Classification of AI systems | OECD Digital Economy Papers. Risk management should be continuous, timely, and performed throughout the AI system lifecycle, starting with the plan & design function in the application context.

The broad audience of the AI RMF is shown in Figure 1. It is composed of AI actors with a variety of roles described below and in Appendix A who must work together to manage the risk and achieve the goals of trustworthy and responsible AI.

The primary audience for using this framework is displayed in the Applications Context, Data & Input, AI Model, and Task & Output dimensions of Figure 1. These individuals and teams manage the design, development, deployment, and acquisition of AI systems and will be driving

AI risk management efforts. The primary audience also includes those with responsibilities to commission or fund an AI system and those who are part of the enterprise management structure governing the AI system 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.

Figure 2: AI actors across the AI lifecycle.

Figure 2 lists representative AI actors across the AI lifecycle. AI actors with expertise to carry out TEVV tasks are especially likely to benefit from the Framework. AI actors with TEVV expertise are integrated throughout the AI lifecycle. TEVV tasks are foundational to risk management, providing knowledge and feedback for AI system management and governance. Performed regularly, TEVV tasks assess the system relative to technical, societal, legal, and ethical standards or norms, as well as monitor and assess risks of emergent properties. As a regular process within an AI lifecycle, TEVV allows for both mid-course remediation and post-hoc risk management and mitigation.

The People & Planet dimension of the AI lifecycle represented in Figure 1 presents an additional AI RMF audience: end-users or affected entities who play an important consultative role to the primary audiences. Their insights and input equip others to analyze context, identify, monitor and manage risks of the AI system by providing formal or quasi-formal norms or guidance. They include trade groups, standards developing organizations, advocacy groups, environmental groups, researchers, and civil society organizations. Their actions can designate boundaries for operation (technical, societal, legal, and ethical). They also promote discussion of the tradeoffs needed to balance societal values and priorities related to civil liberties and rights, equity, and the economy.

A sense of collective responsibility among the many AI actors is essential for AI risk management to be successful.

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. Identifying, mitigating, and minimizing risks and potential harms associated with AI technologies are essential steps towards the development of trustworthy AI systems and their appropriate and responsible use. If a risk management framework can help to effectively address, document, and manage AI risk and negative impacts, it can lead to more trustworthy AI systems.

3.1. Understanding Risk, Impacts, and Harms

In the context of the AI RMF, "risk" refers to the composite measure of an event's probability of occurring and the magnitude (or degree) of the consequences of the corresponding events. The impacts, or consequences, of AI systems can be positive, negative, or both and can result in opportunities or threats (Adapted from: ISO 31000:2018). When considering the negative impact of a potential event, risk is a function of 1) the negative impact, or magnitude of harm, that would arise if the circumstance or event occurs and 2) the likelihood of occurrence (Adapted from: OMB Circular A-130:2016). Negative impact or harm can be experienced by individuals, groups, communities, organizations, society, the environment, and the planet.

While risk management processes address negative impacts, this framework offers approaches to minimize anticipated negative impacts of AI systems *and* identify opportunities to maximize positive impacts. Additionally, the AI RMF is designed to be responsive to new risks as they emerge. This flexibility is particularly important where impacts are not easily foreseeable and applications are evolving. While some AI risks and benefits are well-known, it can be challenging to assess the degree to which a negative impact is related to actual harms. Figure 3 provides examples of potential harms that can be related to AI systems.

> "Risk management refers to coordinated activities to direct and control an organization with regard to risk" (Source: ISO 31000:2018).

Risk management can drive AI developers and users to understand and account for the inherent uncertainties and inaccuracies in their models and systems, which in turn can improve their overall performance and trustworthiness.

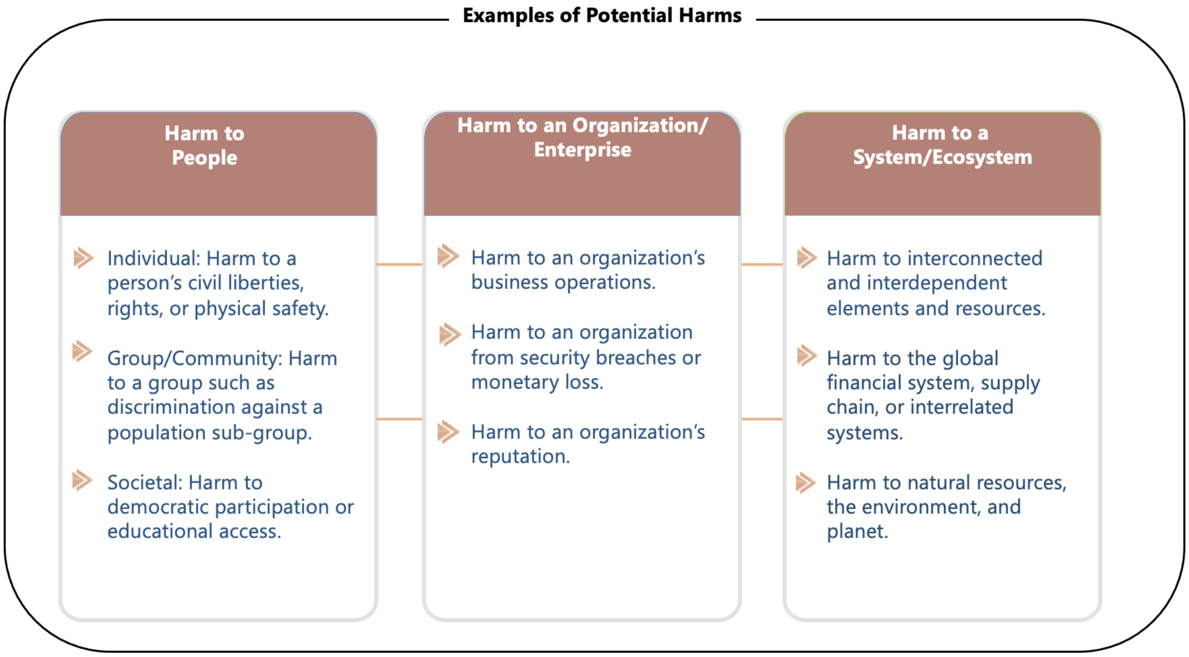

Examples of Potential Harms
Harm to People
Individual: Harm to a person's civil liberties, rights, or physical safety.
Group/Community: Harm to a group such as discrimination against a population sub-group.
Societal: Harm to democratic participation or educational access.
Harm to an Organization/ Enterprise
Harm to an organization's business operations.
Harm to an organization from security breaches or monetary loss.
Harm to an organization's reputation.
Harm to a System/Ecosystem
Harm to interconnected and interdependent elements and resources.
Harm to the global financial system, supply chain, or interrelated systems.
Harm to natural resources, the environment, and planet.

Figure 3: Examples of potential harms related to AI systems. Trustworthy AI systems and their responsible use can mitigate risks and contribute to benefits for people, organizations, and systems.

3.2. Challenges for AI Risk Management

The AI RMF aims to help organizations address a variety of essential risk management issues, emphasizing that context is critical. Several challenges described below stand out for managing risks in pursuit of AI trustworthiness.

3.2.1. Risk Measurement

AI risks and impacts that are not well-defined or adequately understood are difficult to measure quantitatively or qualitatively.

Third-party data or systems can accelerate research and development and facilitate technology transition. They may also complicate risk measurement because the metrics or methodologies used by the organization *developing* the AI system may not align (or may not be transparent or documented) with the metrics or methodologies used by the organization *deploying or operating* the system. Risk measurement and management can further be complicated by how third-party data or systems are used or integrated into AI products or services.

Organizations will want to identify and track emergent risks and consider techniques for measuring them.

Approaches for measuring impacts on a population should consider that harms affect different groups and contexts differently. AI system impact assessments can help AI actors understand potential impacts or harms within specific contexts.

Measuring risk at an earlier stage in the AI lifecycle may yield different results than measuring risk at a later stage. Other risks may be latent at a given time but may increase as AI systems evolve.

AI risks measured in a laboratory or a controlled environment may differ from risks that emerge in operational setting or the real world.

Furthermore, inscrutable AI systems can complicate the measurement of risk. Inscrutability can be a result of the opaque quality of AI systems (lack of explainability or interpretability), lack of transparency or documentation in AI system development or deployment, or inherent uncertainties in AI systems.

3.2.2. Risk Tolerance

While the AI RMF can be used to prioritize risk, it does not prescribe risk tolerance. Risk tolerance refers to the organization's or stakeholder's readiness or appetite to bear the risk in order to achieve its objectives. Risk tolerance can be influenced by legal or regulatory requirements (Adapted from: ISO Guide 73). Risk tolerance and the level of risk that is acceptable to organizations or society are highly contextual and application and use-case specific. Risk tolerances can be influenced by policies and norms established by AI system owners, organizations, industries, communities, or policy makers. Risk tolerances are likely to change and adapt over time as AI systems, policies, and norms evolve. In addition, different organizations may have different risk tolerances due to varying organizational priorities and resource considerations. Even within a single organization there can be a balancing of priorities and tradeoffs.

Emerging knowledge and methods to better inform cost-benefit tradeoffs will continue to be developed and debated by business, governments, academia, and civil society. To the extent that challenges for specifying risk tolerances remain unresolved, there may be contexts where a risk management framework is not yet readily applicable for mitigating AI risks. In the absence of risk tolerances prescribed by existing law, regulation, or norms, the AI RMF equips organizations to define reasonable risk tolerance, manage those risks, and document their risk management process.

When applying the AI RMF, risks which the organization determines to be highest for AI systems and contexts call for the most urgent prioritization and most thorough risk management process. Doing so can mitigate risks and harms to enterprises and individuals, communities, and society. In some cases where an AI system presents the highest risk – where negative impacts are imminent, severe harms are actually occurring, or catastrophic risks are present – development and deployment should cease in a safe manner until risks can be sufficiently mitigated. Conversely, the lowest-risk AI systems and contexts suggest lower prioritization.

3.2.3. Risk Perspectives

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. A risk mitigation culture can help organizations recognize that not all AI risks are the same, so they can allocate resources purposefully. Risk management efforts should align to the risk-level and impact of an AI system, and policies should lay out clear guidelines for assessing each AI system an organization deploys.

3.2.4. Organizational Integration of Risk

The AI RMF is neither a checklist nor a compliance mechanism to be used in isolation. It should be integrated within the organization developing and using AI technologies and incorporated into broader risk management strategy and processes. Thereby AI will be treated along with other critical risks, yielding a more integrated outcome and resulting in organizational efficiencies.

In some scenarios the AI RMF may be utilized along with related guidance and frameworks for managing AI system risks or broader enterprise risks. Some risks related to AI systems are common across other types of software development and deployment. Overlapping risks include privacy concerns related to the use of underlying data to train AI systems, and security concerns related to the confidentiality, integrity and availability of training and output data for AI systems.

Organizations need to establish and maintain the appropriate accountability mechanisms, roles and responsibilities, culture, and incentive structures for risk management to be effective. Use of the AI RMF alone will not lead to these changes or provide the appropriate incentives. Effective risk management needs organizational commitment at senior levels and may require cultural change within an organization or industry. In addition, small to medium-sized organizations may face challenges in implementing the AI RMF which can be different from those of large organizations, depending on their capabilities and resources.

4. AI Risks and Trustworthiness

Approaches which enhance AI trustworthiness can also contribute to a reduction of AI risks. This Framework articulates the following characteristics of trustworthy AI, and offers guidance for addressing them. **Trustworthy AI is: valid and reliable, safe, fair and bias is managed, secure and resilient, accountable and transparent, explainable and interpretable, and privacy-enhanced.**

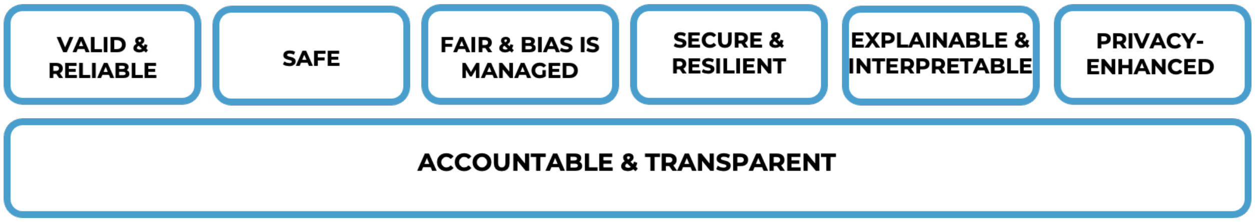


Figure 4: AI trustworthy characteristics.

These characteristics are inextricably tied to human social and organizational behavior, the datasets used by AI systems and the decisions made by those who build them, and the interactions with the humans who provide insight from and oversight of such systems. Human judgment must be employed when deciding on the specific metrics related to AI trustworthy characteristics and the precise threshold values for their related metrics.

Addressing AI trustworthy characteristics individually will not assure AI system trustworthiness, and tradeoffs are always involved. Trustworthiness is greater than the sum of its parts. Ultimately, it is a social concept, and the characteristics listed in any single guidance document will be more or less important in any given situation to establish trustworthiness.

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.

Trustworthiness characteristics explained in this document are interrelated. Highly secure but unfair systems, accurate but opaque and uninterpretable systems, and inaccurate but secure, privacy-enhanced, and transparent systems are all undesirable. Trustworthy AI systems should achieve a high degree of control over risk while retaining a high level of performance quality. Achieving this difficult goal requires a comprehensive approach to risk management, with tradeoffs among the trustworthiness characteristics. It is the joint responsibility of all AI actors to determine whether AI technology is an appropriate or necessary tool for a given context or purpose, and how to use it responsibly. The decision to commission or deploy an AI system should be based on a contextual assessment of trustworthiness characteristics and the relative risks, impacts, costs, and benefits, and informed by a broad set of stakeholders.

Table 1 maps the AI RMF taxonomy to the terminology used by the OECD in their Recommendation on AI, the proposed European Union (EU) Artificial Intelligence Act, and United States Executive Order (EO) 13960.

Table 1: 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

Human Factors

Since AI systems can make sense of information more quickly and consistently than humans, they are often deployed in high-impact settings as a way to make decisions fairer and more impartial than human decision-making, and to do so more efficiently. One common strategy for managing risks in such settings is the use of a human "in-the-loop" (HITL). Unclear expectations about how the HITL can provide oversight for systems, and imprecise governance structures for their configurations are two points for consideration in AI risk management. Identifying which actor should provide which oversight task can be imprecise, with responsibility often falling on the human experts involved in AI-based decision-making tasks. In many settings such experts provide their insights about particular domain knowledge, and are not necessarily able to perform intended oversight or governance functions for AI systems they played no role in developing. It isn't just HITLs; any AI actor, regardless of oversight role, carries their own cognitive biases into the design, development, deployment, and use of AI systems. Biases can be induced by AI actors across the AI lifecycle via assumptions, expectations, and decisions during modeling tasks. These challenges are exacerbated by AI system opacity and the resulting lack of interpretability. The degree to which humans are empowered and incentivized to challenge AI system suggestions must be understood. Data about the frequency and rationale with which humans overrule AI system suggestions in deployed systems can be useful to collect and analyze.

4.1. Valid and Reliable

Accuracy, and robustness are interdependent factors contributing to the validity and trustworthiness of AI systems. 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.

Measures of accuracy – "closeness of results of observations, computations, or estimates to the true values or the values accepted as being true" (Source: ISO/IEC TS 5723:2022) – should address both computational-centric (e.g., false positive and false negative rates) and human-AI teaming aspects. Accuracy measurements should always be paired with clearly defined test sets and details about test methodology; both should be included in associated documentation.

Reliability – "ability of an item to perform as required, without failure, for a given time interval, under given conditions" (Source: ISO/IEC TS 5723:2022) 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.

Robustness or generalizability – "ability of an AI system to maintain its level of performance under a variety of circumstances" (Source: ISO/IEC TS 5723:2022) – is a goal for appropriate system functionality in a broad set of conditions and circumstances, including uses of AI systems not initially anticipated. Robustness does not only require that the system perform exactly as it does under expected uses, but also that it should perform in ways that minimize potential harms to people if it is operating in an unexpected environment.

Validity and reliability for deployed AI systems is often assessed by ongoing audits or monitoring that confirm a system is performing as intended. Measurement of accuracy, reliability, and robustness contribute to trustworthiness and should consider that certain types of failures can cause greater harm – and risks should be managed to minimize the negative impact of those failures.

4.2. Safe

AI systems "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" (Source: ISO/IEC TS 5723:2022). Safe operation of AI systems requires responsible design and development practices, clear information to deployers on how to use a system appropriately, and responsible decision-making by deployers and end-users.

Employing safety considerations during planning and design can prevent failures or conditions that can render a system dangerous. Other practical approaches for AI safety often relate to rigorous simulation and in-domain testing, real-time monitoring, and the ability to shut down or modify systems that deviate from intended or expected functionality.

AI safety measures should take cues from measures of safety used in other fields, such as transportation and healthcare.

4.3. Fair – and Bias Is Managed

Fairness in AI includes concerns for equality and equity by addressing issues such as bias and discrimination. Standards of fairness can be complex and difficult to define because perceptions of fairness differ among cultures and may shift depending on application. Systems in which biases are mitigated are not necessarily fair. For example, systems in which predictions are somewhat balanced across demographic groups may still be inaccessible to individuals with disabilities or affected by the digital divide.

NIST has identified three major categories of AI bias to be considered and managed: systemic, computational, and human, all of which can occur in the absence of prejudice, partiality, or discriminatory intent. Systemic bias can be present in AI datasets, the organizational norms, practices, and processes across the AI lifecycle, and the broader society that uses AI systems. Computational bias can be present in AI datasets and algorithmic processes, and often stems from systematic errors due to non-representative samples. Human biases relate to how an individual or group perceives AI system information to make a decision or fill in missing information. Human biases are omnipresent in decision-making processes across the AI lifecycle and system use. Human biases are implicit, so increasing awareness does not assure control or improvement.

Bias exists in many forms, and can become ingrained in the automated systems that help make decisions about our lives. While bias is not always a negative phenomenon, certain biases exhibited in AI models and systems can perpetuate and amplify negative impacts on individuals, groups, communities, organizations, and society – and at a speed and scale far beyond the traditional discriminatory practices that can result from implicit human or systemic biases. Bias is tightly associated with the concepts of transparency as well as fairness in society. (See NIST Special Publication 1270, "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence.")

4.4. 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 (Adapted from: ISO/IEC TS 5723:2022) may be said to be resilient. AI systems that can maintain confidentiality, integrity, and availability through protection mechanisms that prevent unauthorized access and use may be said to be secure.

Security and resilience are related but distinct characteristics. While resilience is the ability to return to normal function after an attack, security includes resilience but also encompasses protocols to avoid or protect against attacks. Resilience has some relationship to robustness except that it goes beyond the provenance of the data to encompass unexpected or adversarial use of the model or data. Other common security concerns relate to data poisoning and the exfiltration of models, training data, or other intellectual property through AI system endpoints.

4.5. Transparent and Accountable

Transparency 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. Its scope spans from design decisions and training data to model training, the structure of the model, its intended use case, and how and when deployment or end user decisions were made and by whom. Transparency is often necessary for actionable redress related to AI system outputs that are incorrect or otherwise lead to negative impacts. A transparent system is not necessarily an accurate, privacy-enhanced, or secure, or fair system. It is difficult to determine whether an opaque system possesses such characteristics and to do so over time as complex systems evolve.

Determinations of accountability in the AI context relate to expectations of the responsible party in the event that a risky outcome is realized. The shared responsibility of all AI actors should be considered when seeking to hold actors accountable for the outcomes of AI systems. The relationship between risk and accountability associated with AI and technological systems more broadly differs across cultural, legal, sectoral, and societal contexts. Grounding organizational practices and governing structures for harm reduction, like risk management, can help lead to more accountable systems.

Maintaining the provenance of training data and supporting attribution of decisions of the AI system to subsets of training data can assist with both transparency and accountability.

4.6. Explainable and Interpretable

Explainability refers to 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. Together, they assist those operating or overseeing an AI system to do so effectively and responsibly. The underlying assumption is that perceptions of risk stem from a lack of ability to make sense of, or contextualize, system output appropriately.

Risk from lack of explainability may be managed by descriptions of how models work tailored to individual differences such as the user's knowledge and skill level. Explainable systems can be more easily debugged and monitored, and they lend themselves to more thorough documentation, audit, and governance. Risks to interpretability can often be addressed by communicating a description of why an AI system made a particular prediction or recommendation. (See NISTIR 8312, "Four Principles of Explainable Artificial Intelligence" and NISTIR 8367, "Psychological Foundations of Explainability and Interpretability in Artificial Intelligence".)

4.7. Privacy-Enhanced

Privacy refers generally to the norms and practices that help to safeguard human autonomy, identity, and dignity. These norms and practices typically address freedom from intrusion, limiting observation, or individuals' agency to consent to disclosure or control of facets of their identities (e.g., body, data, reputation). (See The NIST Privacy Framework: A Tool for Improving Privacy through Enterprise Risk Management.)

Privacy values such as anonymity, confidentiality, and control generally should guide choices for AI system design, development, and deployment. From a policy perspective, privacy-related risks may overlap with security, bias, and transparency. Like safety and security, specific technical features of an AI system may promote or reduce privacy, and assessors can identify how the processing of data could create privacy-related problems.

5. Effectiveness of the AI RMF

The goal of the AI RMF is to 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. Evaluations of AI RMF effectiveness – including ways to measure bottom-line improvements in the trustworthiness of AI systems – will be part of future NIST activities, in conjunction with stakeholders.

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. The Framework users are expected to benefit from:

» enhanced processes for governing, mapping, measuring, and managing AI risk, and clearly documenting outcomes;
» enhanced awareness of the relationships between and among trustworthiness characteristics, socio-technical approaches, and AI risks;
» established processes for making go/no-go system commissioning and deployment decisions;
» established policies, processes, practices, and procedures for improving organizational accountability efforts related to AI system risks;
» enhanced organizational culture which prioritizes the identification and management of AI system risks and impacts to individuals, communities, organizations, and society;
» enhanced information sharing within and across organizations about decision-making processes, responsibilities, common pitfalls, and approaches for continuous improvement;
» enhanced contextual knowledge for increased awareness of downstream risks;
» enhanced awareness of the importance and efficacy of stakeholder engagement efforts; and
» enhanced capacity for TEVV of AI systems and associated risks.

Part 2: Core and Profiles

6. AI RMF Core

The AI RMF Core provides outcomes and actions that enable dialogue, understanding, and activities to manage AI risks. As illustrated in Figure 5, the Core is composed of four functions: Map, Measure, Manage, and Govern. Each of these high-level functions is broken down into categories and subcategories. Categories and subcategories are subdivided into specific outcomes and actions. Actions do not constitute a checklist, nor are they necessarily an ordered set of steps.

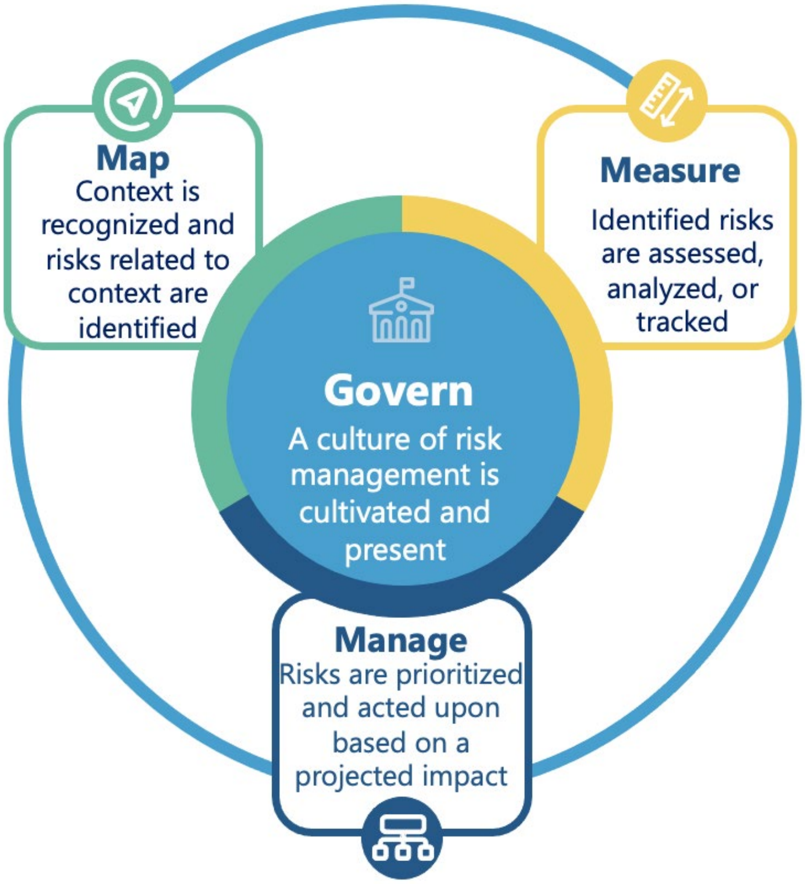

Map
Context is recognized and risks related to context are identified
Measure
Identified risks are assessed, analyzed, or tracked
Govern
A culture of risk management is cultivated and present
Manage
Risks are prioritized and acted upon based on a projected impact

Figure 5: Functions organize AI risk management activities at their highest level to govern, map, measure, and manage AI risks. Governance is a cross-cutting function that is infused throughout and informs the other functions of the process.

Framework users may apply these functions as best suits their needs for managing AI risks. Some organizations may choose to select from among the categories and subcategories; others will want and have the capacity to apply all categories and subcategories. Assuming a governance structure is in place, functions may be performed in any order across the AI lifecycle as deemed to add value by a user of the framework. After instituting the outcomes in Govern, most users of the AI RMF would start with the Map function and continue to Measure or Manage. However users integrate the functions, the process should be iterative, with cross-referencing between functions as necessary. Similarly, there are categories and subcategories with elements that apply to multiple functions, or that have to occur before certain subcategory decisions.

Risk management should be continuous, timely, and performed throughout the AI system lifecycle dimensions. AI RMF core functions should be carried out in a way that reflects diverse and multidisciplinary perspectives, potentially including the views of stakeholders from outside the organization. Having a diverse team contributes to more open sharing of ideas and assumptions about the purpose and function of the technology being designed and developed – which can create opportunities for surfacing problems and identifying existing and emergent risks.

An online companion resource to the AI RMF, referred to as the NIST AI RMF Playbook, is available to help organizations navigate the AI RMF and achieve the outcomes through suggested tactical actions they can apply within their own contexts. The Playbook is voluntary and organizations can utilize the suggestions according to their needs and interests. Playbook users can create tailored guidance from suggested material for their own use, and contribute their suggestions for inclusion to the broader Playbook community. Along with the AI RMF, the NIST Playbook will be part of the forthcoming Trustworthy and Responsible AI Resource Center.

6.1. Govern

The Govern function cultivates and implements a culture of risk management within organizations developing, deploying, or acquiring AI systems. Governance is designed to ensure risks and potential impacts are identified, measured, and managed effectively and consistently. Governance provides a structure through which AI risk management functions can align with organizational policies and strategic priorities whether or not they are related to AI systems.

Governance focuses on technical aspects of AI system design and development as well as on organizational practices and competencies that directly affect the individuals involved in training, deploying, and monitoring of such systems. Governance should address supply chains, including third-party software or hardware systems and data as well as internally developed AI systems.

Govern is a cross-cutting function that is infused throughout AI risk management and informs the other functions of the process. Aspects of Govern, especially those related to compliance or evaluation, should be integrated into each of the other functions. Attention to governance is a continual and intrinsic requirement for effective AI risk management over an AI system's lifespan and the organization's hierarchy.

Senior leadership sets the tone for risk management within an organization, and with it, organizational culture. The governing authorities determine the overarching policies that align with the organization's mission, goals, values, and risk appetite as well as its culture. Management aligns the technical aspects of AI risk management to its policies and operations. Documentation should be used in transparency-enhancing and human review processes, and to bolster accountability in AI system teams.

After putting in place the structures, systems, and teams described in the Govern function, Framework users should be better equipped to carry out meaningful risk management of AI products, services, and systems. It is incumbent on Framework users to continue to execute the Govern function as cultures, stakeholder needs and expectations, and knowledge evolve over time.

Practices related to governing AI risks are described in the NIST AI RMF Playbook.

Table 2 lists the Govern function's categories and subcategories.

Table 2: Categories and subcategories for the Govern function.

Category	**Subcategory**
Govern: A culture of risk management is cultivated and present	
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.

Category	Subcategory
	GOVERN 2.3: Executive leadership of the organization considers decisions about risks associated with AI system development and deployment to be their responsibility.
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.
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.

6.2. Map

The Map function establishes the context to frame risks related to an AI system. The information gathered while carrying out this function enables risk prevention and informs decisions for processes such as model management, and an initial decision about appropriateness or the need for an AI solution. Determination of whether AI use is appropriate or warranted can be considered in comparison to the status quo per a qualitative or quantitative analysis of benefits, costs, and risks. Outcomes in the Map function are the basis for the Measure and Manage functions. Without contextual knowledge, and awareness of risks within the identified contexts, risk management is difficult to perform.

Implementing this function necessitates a broad set of perspectives from a diverse internal team and engagement with external stakeholders. Gathering broad perspectives can help organizations proactively prevent risks and develop more trustworthy AI systems, by

- improving their capacity for understanding contexts;
- checking their assumptions about context of use;
- enabling recognition of when systems are not functional within or out of their intended context;
- identifying positive and beneficial uses of their existing AI systems, and new markets;
- improving understanding of limitations in processes such as proxy development; and
- identifying constraints in real-world applications that may lead to negative impacts.

After completing the Map function, Framework users should have sufficient contextual knowledge about AI system impacts to inform a go/no-go decision about whether to design, develop, or deploy an AI system based on an assessment of impacts. If a decision is made to proceed, organizations should utilize the Measure and Manage functions to assist in AI risk management efforts, utilizing policies and procedures put into place in the Govern function. It is incumbent on Framework users to continue applying the Map function to AI systems as context, capabilities, risks, benefits, and impacts evolve over time.

Practices related to mapping AI risks are described in the NIST AI RMF Playbook.

Table 3 lists the Map function's categories and subcategories.

Table 3: Categories and subcategories for the Map function.

Category	**Subcategory**
Map: Context is recognized and risks related to the context are identified	
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.

Category	**Subcategory**
	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.
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 construct 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.
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.

6.3. Measure

The Measure function employs quantitative, qualitative, or mixed-method tools, techniques, and methodologies to analyze, assess, benchmark, and monitor AI risk and related impacts. It uses knowledge relevant to AI risks identified in the Map function and informs the Manage function. AI systems should be tested before their deployment and regularly while in operation.

Measuring AI risks includes tracking metrics for trustworthy characteristics, social impact, and human-AI configurations. Processes developed or adopted in the Measure function should include rigorous software testing and performance assessment methodologies that include associated measures of uncertainty, comparisons to performance benchmarks, and formalized reporting and documentation of results. Independent review improves the effectiveness of testing and can mitigate internal biases and potential conflicts of interest.

Where tradeoffs among the trustworthy characteristics arise, measurement provides a traceable basis to inform management decisions. Options may include recalibration, impact mitigation, or removal of the system from production.

After completing the Measure function, TEVV processes including metrics, methods, and methodologies are in place, followed, and documented. Framework users will enhance their capacity to comprehensively evaluate system trustworthiness, identify and track existing and emergent risks, and verify efficacy of metrics. Measurement outcomes will be utilized in the Manage function to assist risk monitoring and response efforts. It is incumbent on Framework users to continue applying the Measure function to AI systems as knowledge, methodologies, risks, and impacts evolve over time.

Practices related to measuring AI risks will be described in the NIST AI RMF Playbook.

Table 4 lists the Measure function's categories and subcategories.

Table 4: Categories and subcategories for the Measure function.

Category	Subcategory
Measure: Identified risks are assessed, analyzed, or tracked	
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.
	MEASURE 2.1: Test sets, metrics, and details about the tools used during test, evaluation, validation, and verification (TEVV) are documented.

Category	Subcategory
MEASURE 2: Systems are evaluated for trustworthy characteristics.	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.
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.

6.4. Manage

The Manage function entails allocating risk management resources to mapped and measured risks on a regular basis and as defined by the Govern function.

Contextual information gleaned from stakeholder feedback and other expert consultation processes established in Govern and carried out in Map are also utilized in this function to decrease the likelihood of system failures and negative impacts. Systematic documentation practices established in Govern and utilized in Map and Measure bolster AI risk management efforts and increase transparency and accountability.

After completing the Manage function, plans for prioritizing risk and continuous monitoring and improvement will be in place. Framework users will have enhanced capacity to manage the risks of deployed AI systems and to allocate risk management resources based on risk measures. It is incumbent on Framework users to continue to apply the Manage function to deployed AI systems as methods, contexts, risks, and stakeholder expectations evolve over time.

Practices related to managing AI risks will be described in the NIST AI RMF Playbook.

Table 5 lists the Manage function's categories and subcategories.

Table 5: Categories and subcategories for the Manage function.

Category	**Subcategory**
Manage: Risks are prioritized and acted upon based on a projected impact	
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.
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.

Category	**Subcategory**
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.

7. AI RMF Profiles

AI RMF *use case profiles* are instantiations of the AI RMF functions, categories, and subcategories for a certain application or use case based on the requirements, risk tolerance, and resources of the Framework user. Examples could be an AI RMF *hiring profile* or an AI RMF *fair housing profile*. Profiles may illustrate and offer insights into how risk can be managed at various stages of the AI lifecycle or in specific sector, technology, or end-use applications. A profile assists organizations in deciding how they might best manage AI risk that is well-aligned with their goals, considers legal/regulatory requirements and best practices, and reflects risk management priorities.

AI RMF *temporal profiles* are descriptions of either the current state or the desired, target state of specific AI risk management activities within a given sector, industry, organization, or application context. An AI RMF *Current Profile* indicates how AI is currently being managed and the related risks in terms of current outcomes. A *Target Profile* indicates the outcomes needed to achieve the desired or target AI risk management goals.

Comparing Current and Target Profiles may reveal gaps to be addressed to meet AI risk management objectives. Action plans can be developed to address these gaps to fulfill a given Category or Subcategory. Prioritizing the mitigation of gaps is driven by the user's needs and risk management processes. This risk-based approach enables Framework users to compare their approaches and themselves with other stakeholders and to gauge the resources needed (e.g., staffing, funding) to achieve AI risk management goals in a cost-effective, prioritized manner.

This Framework does not prescribe Profile templates, allowing for flexibility in implementation.

NOTE: NIST welcomes contributions towards development of AI RMF use case profiles as well as current and target profiles. Submissions to be included in the NIST Trustworthy and Responsible AI Resource Center will inform NIST and the broader community about the usefulness of the AI RMF and will likely lead to improvements which can be incorporated into future versions of the framework.

Appendix A: Descriptions of AI Actor Tasks from Figure 1

AI Design includes AI actors who are responsible for the planning, design, and data collection and processing tasks of the AI system. Tasks include articulating and documenting the system's concept and objectives, underlying assumptions, context, and requirements; gathering and cleaning data; and documenting the metadata and characteristics of the dataset. AI actors in this category include data scientists, domain experts, socio-cultural analysts, human factors experts, governance experts, data engineers, data providers, and evaluators.

AI Development includes AI actors who are responsible for model building and interpretation tasks, which involve the creation, selection, calibration, training, and/or testing of models or algorithms. Tasks involve machine learning experts, data scientists, developers, and experts with familiarity about the socio-cultural and contextual factors associated with the deployment setting.

AI Deployment includes AI actors who assure deployment of the system into production. Related tasks include: piloting, checking compatibility with legacy systems, ensuring regulatory compliance, managing organizational change, and evaluating user experience. AI actors in this category include system integrators, software developers, evaluators and domain experts with expertise in human factors, socio-cultural analysis, and governance.

Operation and Monitoring includes AI actors who are responsible for operating the AI system and working with others to continuously assess system output and impacts. Users who interpret or incorporate the output of AI systems, evaluators and auditors, and members of the research community are part of this group.

Test, Evaluation, Verification, and Validation (TEVV) tasks are performed by AI actors who examine the AI system or its components, or detect and remediate problems throughout the AI lifecycle. Tasks can be incorporated into a phase as early as design, where tests are planned in accordance with the design requirement.

- TEVV tasks for design, planning, and data may center on internal and external validation of assumptions for system design, data collection, and measurements, relative to the intended context of deployment or application.
- TEVV tasks for development (i.e., model building) include model validation and assessment.
- TEVV tasks for deployment include system validation and integration in production, with testing, tuning, and recalibration for systems and process integration, user experience, and compliance with existing legal, regulatory, and ethical specifications.
- TEVV tasks for operations involve ongoing monitoring for periodic updates, testing, and recalibration of models, and the detection of emergent properties and related impacts.

Human Factors tasks and activities include human-centered design practices and methodologies, promoting the active involvement of end-users and appropriate stakeholders, incorporating context-specific norms and values in system design (VSD), evaluating and adapting end-user experiences, and broad integration of humans and human dynamics in all phases of the AI lifecycle. Human factors professionals provide multidisciplinary skills and perspectives to understand context of use, engage multi-stakeholder processes, design and evaluate user experience (UI/UX), perform human-centered evaluation and testing, and inform impact assessments.

Domain Experts are multidisciplinary practitioners or scholars who provide knowledge or expertise in an industry sector, economic sector, or application area where an AI system is being used. These experts are essential contributors for AI system design and development and can provide interpretation of outputs to support the work of TEVV and AI impact assessment teams.

AI Impact Assessors are responsible for assessing and evaluating requirements for AI system accountability, combating harmful bias, examining intended and unintended impacts of AI systems, product safety, liability, and security, among others. AI Impact assessors provide technical, human factor, socio-cultural, and legal expertise.

Procurers are financial, legal, or policy management officials who acquire AI models, products, or services from a third party, developer, vendor, or contractor.

Third-party entities are providers, developers, or vendors of data, algorithms, models, and/or systems and related services to another organization or the organization's customers or clients. Third-party entities are responsible for AI design and development tasks, in whole or in part. By definition, they are external to the design, development, or deployment team of the organization that acquires its technologies or services. The technologies acquired from third-party entities may be complex or opaque, and risk tolerances may not align with the deploying or operating organization.

Organizational Management, Senior Leadership, and **the Board of Directors** are among the parties responsible for AI governance.

End Users of an AI system are the individuals or groups that use the system for a specific purpose. These individuals or groups interact with an AI system in a specific context. End users can range in competency from AI experts to first-time technology end-users.

AI Operators continuously assess system recommendations and impacts (both intended and unintended) in light of the system's objectives as well as the ethical considerations that go into its operation. Operators can often be associated with the planning and design or specification stage of the AI system lifecycle as well as post-deployment monitoring.

Affected Individuals/Communities encompass any individual, group, community, or stakeholder organization affected by AI systems or decisions based on the output of AI systems, directly or indirectly. These individuals do not necessarily interact with the system and can be indirectly or directly affected by the deployment of an AI system or application.

Other AI actors may provide formal or quasi-formal norms or guidance for specifying and managing AI risks. They can include **trade groups**, **standards developing organizations**, **advocacy groups, environmental groups,** and **civil society organizations**.

The **general public** is most likely to directly experience positive and negative impacts of AI technologies. They may provide the motivation for actions taken by the other stakeholders and can include individuals, communities, and consumers in the context where an AI system is developed or deployed.

Appendix B: How AI Risks Differ from Traditional Software Risks

As with traditional software, risks from AI-based technology can be bigger than an enterprise, span organizations, and can lead to societal impacts. AI systems also bring a set of risks that are not comprehensively addressed by current risk frameworks and approaches. Some AI systems' features that present risks can also be beneficial. For example, pre-trained models and transfer learning can advance research and increase accuracy and resilience when compared to other models and approaches. Identifying contextual factors in the Map function will assist AI actors in determining the level of risk and potential management efforts.

Compared to traditional software, AI-specific risks that are new or increased include:

- "Oracle problem" - the data used for building an AI system is considered oracle, but data may not be a true or appropriate representation of the context or intended use of the AI system. Additionally, bias and other data quality issues can affect AI system trustworthiness, which could lead to negative impacts.
- AI system dependency and reliance on data for training tasks, combined with increased volume and complexity typically associated with such data.
- Intentional or unintentional changes during training that may fundamentally alter AI system performance.
- Datasets used to train AI systems may become detached from their original and intended context, or may become stale or outdated relative to deployment context.
- AI system scale and complexity (many systems contain billions or even trillions of decision points) housed within more traditional software applications.
- Use of pre-trained models that can advance research and improve performance can also increase levels of statistical uncertainty and cause issues with bias management, scientific validity, and reproducibility.
- Higher degree of difficulty in predicting failure modes for emergent properties of large-scale pre-trained models.
- Increased opacity and concerns about reproducibility.
- Underdeveloped software testing standards.
- Computational costs for developing AI systems and their impact on the environment and planet.

Current standard privacy and security controls are not able to comprehensively address many of these AI system risks. Existing privacy, computer security, and data security frameworks and guidance are unable to:

- adequately manage the problem of bias in AI systems;
- comprehensively address security concerns related to evasion, model extraction, membership inference, or other machine learning attacks;

» address the complex attack surface of AI systems or other security abuses enabled by AI systems; and

» address risks associated with third-party AI technologies, transfer learning, and off-label use, where AI systems may be trained for decision-making outside an organization's security controls or trained in one domain and then "fine-tuned" for another.

Perceptions about AI system capabilities can be another source of risk. One major false perception is the presumption that AI systems work – and work well – in all settings. Whether accurate or not, AI is often portrayed in public discourse as more objective than humans, and with greater capabilities than general software. Additionally, since systemic biases can be encoded in AI system training data and individual and group decision making across the AI lifecycle, many of the negative system impacts can be concentrated on historically excluded groups.