AI Risk Management Framework: Initial Draft

March 17, 2022

This initial draft of the Artificial Intelligence Risk Management Framework (AI RMF, or Framework) builds on the concept paper released in December 2021 and incorporates the feedback received. The AI RMF is intended for voluntary use in addressing risks in the design, development, use, and evaluation of AI products, services, and systems.

AI research and deployment is evolving rapidly. For that reason, the AI RMF and its companion documents will evolve over time. When AI RMF 1.0 is issued in January 2023, NIST, working with stakeholders, intends to have built out the remaining sections to reflect new knowledge, awareness, and practices.

Part I of the AI RMF sets the stage for why the AI RMF is important and explains its intended use and audience. Part II includes the AI RMF Core and Profiles. Part III includes a companion Practice Guide to assist in adopting the AI RMF.

That Practice Guide which will be released for comment includes additional examples and practices that can assist in using the AI RMF. The Guide will be part of a NIST AI Resource Center that is being established.

NIST welcomes feedback on this initial draft and the related Practice Guide to inform further development of the AI RMF. Comments may be provided at a workshop on March 29-31, 2022, and also are strongly encouraged to be shared via email. NIST will produce a second draft for comment, as well as host a third workshop, before publishing AI RMF 1.0 in January 2023. **Please send comments on this initial draft to AIframework@nist.gov by April 29, 2022**.

Comments are especially requested on:

1. Whether the AI RMF appropriately covers and addresses AI risks, including with the right level of specificity for various use cases.
2. Whether the AI RMF is flexible enough to serve as a continuing resource considering evolving technology and standards landscape.
3. Whether the AI RMF enables decisions about how an organization can increase understanding of, communication about, and efforts to manage AI risks.
4. Whether the functions, categories, and subcategories are complete, appropriate, and clearly stated.
5. Whether the AI RMF is in alignment with or leverages other frameworks and standards such as those developed or being developed by IEEE or ISO/IEC SC42.
6. Whether the AI RMF is in alignment with existing practices, and broader risk management practices.
7. What might be missing from the AI RMF.
8. Whether the soon to be published draft companion document citing AI risk management practices is useful as a complementary resource and what practices or standards should be added.
9. Others?

Note: This first draft does not include Implementation Tiers as considered in the concept paper. Implementation Tiers may be added later if stakeholders consider them to be a helpful feature in the AI RMF. Comments are welcome.

Table of Contents

AI Risk Management Framework: Initial Draft - Part 1: Motivation

1 Overview

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 – everything from commerce and healthcare to transportation and cybersecurity. AI systems are used for tasks such as informing and advising people and taking actions where they can have beneficial impact, such as safety and housing.

AI systems sometimes do not operate as intended because they are making inferences from patterns observed in data rather than a true understanding of what causes those patterns. Ensuring that these inferences are helpful and not harmful in particular use cases – especially when inferences are rapidly scaled and amplified – is fundamental to trustworthy AI. While answers to the question of what makes an AI technology trustworthy differ, there are certain key characteristics which support trustworthiness, including accuracy, explainability and interpretability, privacy, reliability, robustness, safety, security (resilience) and mitigation of harmful bias. There also are key guiding principles to take into account such as accountability, fairness, and equity.

Cultivating trust and communication about how to understand and manage the risks of AI systems will help create opportunities for innovation and realize the full potential of this technology.

Many activities related to managing risk for AI are common to managing risk for other types of technology. An AI Risk Management Framework (AI RMF, or Framework) can address challenges unique to AI systems. This AI RMF is an initial attempt to describe how the risks from AI-based systems differ from other domains and to encourage and equip many different stakeholders in AI to address those risks purposefully.

It is important to note that the AI RMF is neither a checklist nor should be used in any way to certify an AI system. Likewise, using the AI RMF does not substitute for due diligence and judgment by organizations and individuals in deciding whether to design, develop, and deploy AI technologies – and if so, under what conditions.

This voluntary framework provides a flexible, structured, and measurable process to address AI risks throughout the AI lifecycle, offering guidance for the development and use of trustworthy and responsible AI. It is intended to improve understanding of and to help organizations manage both enterprise and societal risks related to the development, deployment, and use of AI systems. Adopting the AI RMF can assist organizations, industries, and society to understand and determine their acceptable levels of risk.

In addition, it can be used to map compliance considerations beyond those addressed by this framework, including existing regulations, laws, or other mandatory guidance.

Risks to any software or information-based system apply to AI; that includes important concerns related to cybersecurity, privacy, safety, and infrastructure. This framework aims to fill the gaps related specifically to AI. Rather than repeat information in other guidance, users of the AI RMF are encouraged to address those non-AI specific issues via guidance already available.

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. Part 3 [yet to be developed] assists in using the AI RMF and offers sample practices to be considered in carrying out this guidance, before, during, and after AI products, services, and systems are developed and deployed.

For the purposes of the NIST AI RMF the term *artificial intelligence* refers to algorithmic processes that learn from data in an automated or semi-automated manner.

The Framework, and supporting resources, will be updated and improved based on evolving technology and the standards landscape around the globe. In addition, as the AI RMF is put into use, additional lessons will be learned that can inform future updates and additional resources.

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

2 Scope

The NIST AI RMF offers a process for managing risks related to AI systems across a wide spectrum of types, applications, and maturity. This framework is organized and intended to be understood and used by individuals and organizations, regardless of sector, size, or level of familiarity with a specific type of technology. Ultimately, it will be offered in multiple formats, including online versions, to provide maximum flexibility.

The AI RMF serves as a part of a broader NIST resource center containing documents, taxonomy, suggested toolkits, datasets, code, and other forms of technical guidance related to the development and implementation of trustworthy AI. Resources will include a knowledge base of terminology related to trustworthy and responsible AI and how those terms are used by different stakeholders.

The AI RMF is not a checklist nor a compliance mechanism to be used in isolation. It should be integrated within the organization developing and using AI and be incorporated into enterprise

risk management; doing so ensures that AI will be treated along with other critical risks, yielding a more integrated outcome and resulting in organizational efficiencies.

Attributes of the AI RMF

The AI RMF strives to:

1. Be risk-based, resource efficient, 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 mesh 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 both technology agnostic and applicable 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.

3 Audience

AI risk management is a complex and relatively new area, and the list of individuals, groups, communities, and organizations that can be affected by AI technologies is extensive. Identifying and managing AI risks and impacts – both positive and adverse – requires a broad set of perspectives and stakeholders.

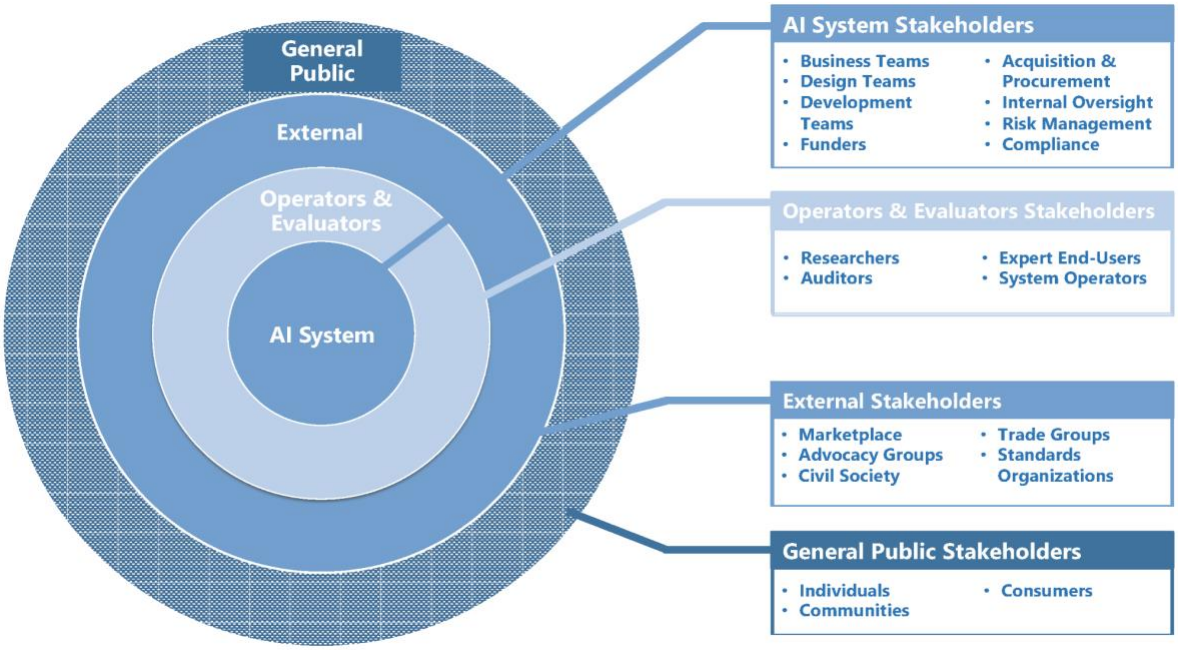

Figure 1: Key stakeholder groups associated with the AI RMF.

As Figure 1 illustrates, NIST has identified four stakeholder groups as intended audiences of this Framework: AI system stakeholders, operators and evaluators, external stakeholders, and the general public. Ideally, members of all stakeholder groups would be involved or represented in the risk management process, including those individuals and community representatives that may be affected by the use of AI technologies.

AI system stakeholders are those who have the most control and responsibility over the design, development, deployment, and acquisition of AI systems, and the implementation of AI risk management practices. This group comprises the primary adopters of the AI RMF. They may include individuals or teams within or among organizations with responsibilities to commission, fund, procure, develop, or deploy an AI system: business teams, design and development teams, internal risk management teams, and compliance teams. Small to medium-sized organizations face different challenges in implementing the AI RMF than large organizations.

Operators and evaluators provide monitoring and formal/informal test, evaluation, validation, and verification (TEVV) of system performance, relative to both technical and socio-technical requirements. These stakeholders, which include organizations which operate or employ AI systems, use the output for decisions or to evaluate their performance. This group can include users who interpret or incorporate the output of AI systems in settings with a high potential for adverse impacts. They might include academic, public, and private sector researchers; professional evaluators and auditors; system operators; and expert end users.

External stakeholders provide formal and/or quasi-formal norms or guidance for specifying and addressing AI risks. External to the primary adopters of the AI RMF, they can include trade

groups, standards developing organizations, advocacy groups, and civil society organizations. Their actions can designate boundaries for operation (technical or legal) and balance societal values and priorities related to civil liberties and rights, the economy, and security.

The **general public** is most likely to directly experience positive and adverse 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.

4 Framing Risk

AI systems hold the potential to advance our quality of life and lead to new services, support, and efficiencies for people, organizations, markets, and society. Identifying, mitigating, and minimizing risks and potential harms associated with AI technologies are essential steps towards the acceptance and widespread use of AI technologies. A risk management framework should provide a structured, yet flexible, approach for managing enterprise and societal risk resulting from the incorporation of AI systems into products, processes, organizations, systems, and societies. Organizations managing an enterprise's AI risk also should be mindful of larger societal AI considerations and risks. If a risk management framework can help to effectively address and manage AI risk and adverse impacts, it can lead to more trustworthy AI systems.

4.1 Understanding Risk and Adverse Impacts

Risk is a measure of the extent to which an entity is negatively influenced by a potential circumstance or event. Typically, risk is a function of 1) the adverse impacts that could arise if the circumstance or event occurs; and 2) the likelihood of occurrence. Entities can be individuals, groups, or communities as well as systems, processes, or organizations.

The impact of AI systems can be positive, negative, or both and can address, create, or result in opportunities or threats. According to the International Organization for Standardization (Guide 73:2009; IEC/ISO 31010), certain risks can be positive. While risk management processes address adverse impacts, this framework intends to offer approaches to minimize anticipated negative impacts of AI systems *and* identify opportunities to maximize positive impacts. Additionally, this framework is designed to be responsive to new risks as they emerge rather than enumerating all known risks in advance. This flexibility is particularly important where impacts are not easily foreseeable, and applications are evolving rapidly. While AI benefits and some AI risks are well-known, the AI community is only beginning to understand and classify incidents and scenarios that result in harm. Figure 2 provides examples of potential harms from AI systems.

Risk management can also drive AI developers and users to understand and account for the inherent uncertainties and inaccuracy of their models and systems, which in turn can increase the

overall performance and trustworthiness of those models. Managing risk and adverse impacts contributes to building trustworthy AI technologies and applications

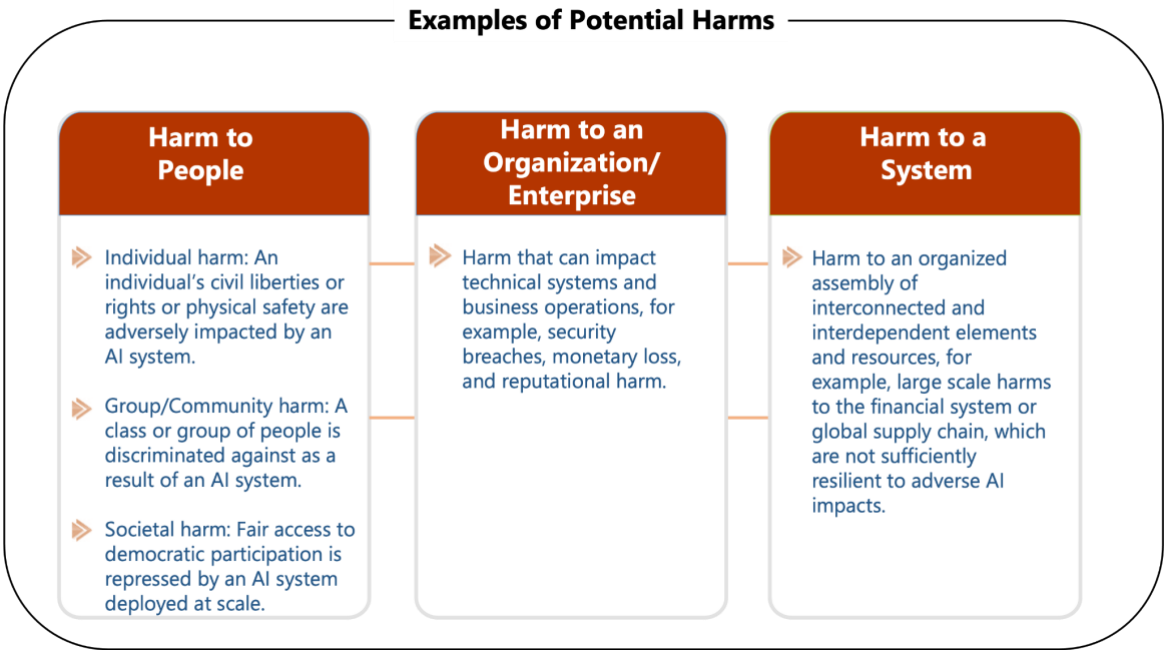

Figure 2: Examples of potential harms from AI systems.

4.2 Challenges for AI Risk Management

4.2.1 Risk Measurement

AI risks and impacts that are not well-defined or adequately understood are difficult to measure quantitatively or qualitatively. The presence of third-party data or systems may also complicate risk measurement. Those attempting to measure the adverse impact on a population may not be aware that certain demographics may experience harm differently than others.

AI risks can have a temporal dimension. Measuring risk at an earlier stage in the AI lifecycle may yield different results than measuring risk at a later stage. Some AI risks may have a low probability in the short term but have a high likelihood for adverse impacts. Other risks may be latent at present but may increase in the long term as AI systems evolve.

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

4.2.2 Risk Thresholds

Thresholds refer to the values used to establish concrete decision points and operational limits that trigger a response, action, or escalation. AI risk thresholds (sometimes referred to as Key Risk Indicators) can involve both technical factors (such as error rates for determining bias) and human values (such as social or legal norms for appropriate levels of transparency). These

factors and values can establish levels of risk (e.g., low, medium, or high) based on broad categories of adverse impacts or harms.

Thresholds and values can also determine where AI systems present unacceptable risks to certain organizations, systems, social domains, or demographics. In these cases, the question is not how to better manage risk of AI, but whether an AI system should be designed, developed, or deployed at all.

The AI RMF does not prescribe risk thresholds or values. Risk tolerance – the level of risk or degree of uncertainty that is acceptable to organizations or society – is context and use case-specific. Therefore, risk thresholds should be set through policies and norms that can be established by AI system owners, organizations, industries, communities, or regulators (who often are acting on behalf of individuals or societies). Risk thresholds and values are likely to change and adapt over time as policies and norms change or evolve. In addition, different organizations may have different risk thresholds (or tolerance) due to varying organizational priorities and resource considerations. Even within a single organization there can be a balancing of priorities and tradeoffs between technical factors and human values. Emerging knowledge and methods for better informing these decisions are being developed and debated by business, governments, academia, and civil society. To the extent that challenges for specifying risk thresholds or determining values remain unresolved, there may be contexts where a risk management framework is not yet readily applicable for mitigating AI risks and adverse impacts. The AI RMF provides the opportunity for organizations to specifically define their risk thresholds and then to manage those risks within their tolerances.

4.2.3 Organizational Integration

The AI RMF is not a checklist nor a compliance mechanism to be used in isolation. It should be integrated within the organization developing and using AI technologies and be incorporated into enterprise risk management; doing so ensures that AI will be treated along with other critical risks, yielding a more integrated outcome and resulting in organizational efficiencies.

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 significant cultural change for an organization or industry.

Small to medium-sized organizations face different challenges in implementing the AI RMF than large organizations.

5 AI Risks and Trustworthiness

The AI RMF uses a three-class taxonomy, depicted in Figure 3, to classify characteristics that should be considered in comprehensive approaches for identifying and managing risk related to AI systems: *technical characteristics, socio-technical characteristics, and guiding principles*.

This AI RMF taxonomy frames AI risk using characteristics that are aligned with trustworthy AI systems, in conjunction with contextual norms and values. Since AI trustworthiness and risk are inversely related, approaches which enhance trustworthiness can contribute to a reduction or attenuation of related risks. The AI RMF taxonomy articulates several key building blocks of trustworthy AI within each category, which are particularly suited to the examination of potential risk.

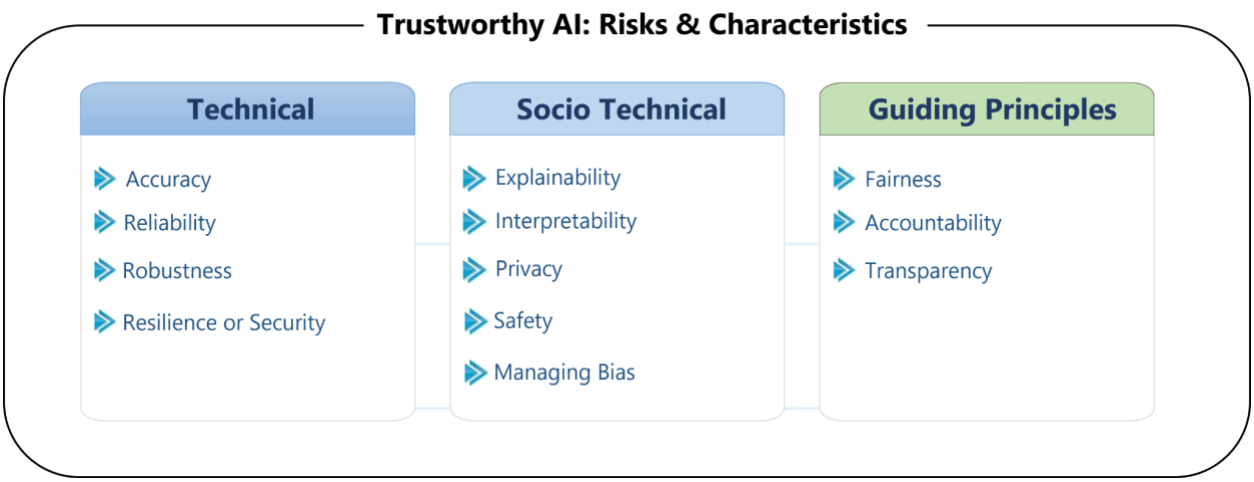

Figure 3: AI Risks and Trustworthiness. The three-class taxonomy to classify characteristics that should be considered in comprehensive approaches for identifying and managing risk related to AI systems. The taxonomy articulates several key building blocks of trustworthy AI within each category, which are particularly suited to the examination of potential risk.

Figure 4 provides a mapping of the AI RMF taxonomy to the terminology used by the Organisation for Economic Co-operation and Development (OECD) in their Recommendation on AI, the European Union (EU) Artificial Intelligence Act, and United States Executive Order (EO) 13960.

5.1 Technical Characteristics

Technical characteristics in the AI RMF taxonomy refer to factors that are under the direct control of AI system designers and developers, and which may be measured using standard evaluation criteria. Technical characteristics include the tradeoff between convergent-discriminant validity (whether the data reflects what the user intends to measure and not other things) and statistical reliability (whether the data may be subject to high levels of statistical noise and measurement bias). Validity of AI, especially machine learning (ML) models, can be assessed using technical characteristics. Validity for deployed AI systems is often assessed with ongoing audits or monitoring that confirm that a system behaves as intended. It may be possible to utilize and automate explicit measures based on variations of standard statistical or ML techniques and specify thresholds in requirements. Data generated from experiments that are designed to evaluate system performance also fall into this category and might include tests of causal hypotheses and assessments of robustness to adversarial attack.

	Technical Design Characteristics	Socio-Technical Characteristics	Guiding Principles Contributing to Trustworthiness
AI RMF Taxonomy	• Accuracy • Reliability • Robustness • Resilience or ML Security	• Explainability • Interpretability • Privacy • Safety • Managing Bias	• Fairness • Accountability • Transparency
OECD AI Recommendation	• Robustness • Security	• Safety • Explainability	• Traceability to human values • Transparency and responsible disclosure • Accountability
EU AI Act	• Technical robustness	• Safety • Privacy • Non-discrimination	• Human agency and oversight • Data governance • Transparency • Diversity and fairness • Environmental and societal well-being • Accountability
EO 13960	• Purposeful and performance-driven • Accurate, reliable, and effective • Secure and resilient	• Safe • Understandable by subject matter experts, users, and others, as appropriate	• Lawful and respectful of our Nation's values • Responsible and traceable • Regularly monitored • Transparent • Accountable

Figure 4: Mapping of AI RMF taxonomy to AI policy documents.

The following technical characteristics lend themselves well to addressing AI risk: accuracy, reliability, robustness, and resilience (or ML security).

5.1.1 Accuracy

Accuracy indicates the degree to which the ML model is correctly capturing a relationship that exists within training data. Analogous to statistical conclusion validity, accuracy is examined via standard ML metrics (e.g., false positive and false negative rates, F1-score, precision, and recall), as well as assessment of model underfit or overfit (high testing errors irrespective of error rates in training). It is widely acknowledged that current ML methods cannot guarantee that the underlying model is capturing a causal relationship. Establishing internal (causal) validity in ML models is an active area of research. AI risk management processes should take into account the potential risks to the enterprise and society if the underlying causal relationship inferred by a model is not valid, calling into question decisions made on the basis of the model. Determining a threshold for accuracy that corresponds with acceptable risk is fundamental to AI risk management and highly context-dependent.

5.1.2 Reliability

Reliability indicates whether a model consistently generates the same results, within the bounds of acceptable statistical error. Techniques designed to mitigate overfitting (e.g., regularization) and to adequately conduct model selection in the face of the bias/variance tradeoff can increase model reliability. The definition of reliability is analogous to construct reliability in the social sciences, albeit without explicit reference to a theoretical construct. Reliability measures may give insight into the risks related to decontextualization, due to the common practice of reusing

ML datasets or models in ways that cause them to become disconnected from the social contexts and time periods of their creation. As with accuracy, reliability provides an evaluation of the validity of models, and thus can be a factor in determining thresholds for acceptable risk.

5.1.3 Robustness

Robustness is a measure of model sensitivity, indicating whether the model has minimum sensitivity to variations in uncontrollable factors. A robust model will continue to function despite the existence of faults in its components. The performance of the model may be diminished or otherwise altered until the faults are corrected. Measures of robustness might range from sensitivity of a model's outputs to small changes in its inputs, but might also include error measurements on novel datasets. Robustness contributes to sensitivity analysis in the AI risk management process.

5.1.4 Resilience or ML Security

A model that can withstand adversarial attacks, or more generally, unexpected changes in its environment or use, may be said to be resilient or secure. This attribute 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 ML security concerns relate to the exfiltration of models, training data, or other intellectual property through AI system endpoints.

5.2 Socio-Technical Characteristics

Socio-technical characteristics in the AI RMF taxonomy refer to how AI systems are used and perceived in individual, group, and societal contexts. This includes mental representations of models, whether the output provided is sufficient to evaluate compliance (transparency), whether model operations can be easily understood (explainability), whether they provide output that can be used to make a meaningful decision (interpretability), and whether the outputs are aligned with societal values. Socio-technical factors are inextricably tied to human social and organizational behavior, from the datasets used by ML processes and the decisions made by those who build them, to the interactions with the humans who provide the insight and oversight to make such systems actionable.

Unlike technical characteristics, socio-technical characteristics require significant human input and cannot yet be measured through an automated process. Human judgment must be employed when deciding on the specific metrics and the precise threshold values for these metrics. The connection between human perceptions and interpretations, societal values, and enterprise and societal risk is a key component of the kinds of cultural and organizational factors that will be necessary to properly manage AI risks. Indeed, input from a broad and diverse set of stakeholders is required throughout the AI lifecycle to ensure that risks arising in social contexts are managed appropriately.

The following socio-technical characteristics have implications for addressing AI risk: explainability, interpretability, privacy, safety, and managing bias.

5.2.1 Explainability

Explainability seeks to provide a programmatic, sometimes causal, description of how model predictions are generated. Even given all the information required to make a model fully transparent, a human must apply technical expertise if they want to understand how the model works. Explainability refers to the user's perception of how the model works – such as what output may be expected for a given input. Explanation techniques tend to summarize or visualize model behavior or predictions for technical audiences. Explanations can be useful in promoting human learning from machine learning, for addressing transparency requirements, or for debugging issues with AI systems and training data. However, risks due to explainability may arise for many reasons, including, for example, a lack of fidelity or consistency in explanation methodologies, or if humans incorrectly infer a model's operation, or the model is not operating as expected. Risk from lack of explainability may be managed by descriptions of how models work to users' skill levels. Explainable systems can be more easily debugged and monitored, and lend themselves to more thorough documentation, audit, and governance.

Explainability is related to transparency. Typically the more opaque a model is, the less it is considered explainable. However, transparency does not guarantee explainability, especially if the user lacks an understanding of ML technical principles.

5.2.2 Interpretability

Interpretability seeks to fill a meaning deficit. Although explainability and interpretability are often used interchangeably, explainability refers to a representation of the mechanisms underlying an algorithm's operation, whereas interpretability refers to the meaning of its output in the context of its designed functional purpose. The underlying assumption is that perceptions of risk stem from a lack of ability to make sense of, or contextualize, model output appropriately. Model interpretability refers to the extent to which a user can determine adherence to this function and the consequent implications of this output upon other consequential decisions for that user. Interpretations are typically contextualized in terms of values and reflect simple, categorical distinctions. For example, a society may value privacy and safety, but individuals may have different determinations of safety thresholds. Risks to interpretability can often be addressed by communicating the interpretation intended by model designers, although this remains an open area of research. The prevalence of different interpretations can be readily measured with psychometric instruments.

5.2.3 Privacy

Privacy refers generally to the norms and practices that help to safeguard values such as human autonomy and dignity. These norms and practices typically address freedom from intrusion, limiting observation, or individuals' control of facets of their identities (e.g., body, data, reputation). Like safety and security, specific technical features of an AI system may promote privacy, and assessors can identify how the processing of data could create privacy-related problems. However, determinations of likelihood and severity of impact of these problems are contextual and vary among cultures and individuals.

5.2.4 Safety

Safety as a concept is highly correlated with risk and generally denotes an absence (or minimization) of failures or conditions that render a system dangerous. As AI systems interact with humans more directly in factories and on the roads, for example, the safety of these systems is a serious consideration for AI risk management. Safety is often – though not always – considered through a legal lens. Practical approaches for AI safety often relate to rigorous simulation and in-domain testing, real-time monitoring, and the ability to quickly shut down or modify misbehaving systems.

5.2.5 Managing Bias

NIST has identified three major categories of bias in AI: systemic, computational, and human. Managing bias in AI systems requires an approach that considers all three categories.
Bias exists in many forms, is omnipresent in society, 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, 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 and fairness in society. See NIST publication “Towards a Standard for Identifying and Managing Bias in Artificial Intelligence.”

> When managing risks in AI systems it is important to understand that the attributes of the AI RMF risk taxonomy are interrelated. Highly secure but unfair systems, accurate but opaque and uninterpretable systems, and inaccurate, but fair, secure, privacy-protected, and transparent systems are all undesirable. It is possible for trustworthy AI systems to achieve a high degree of risk control while retaining a high level of performance quality. Achieving this difficult goal requires a comprehensive approach to risk management, with tradeoffs among the technical and socio-technical characteristics.

5.3 Guiding Principles

Guiding principles in the AI RMF taxonomy refer to broader societal norms and values that indicate societal priorities. While there is no objective standard for ethical values, as they are grounded in the norms and legal expectations of specific societies or cultures, it is widely agreed that AI technologies should be developed and deployed in ways that meet contextual norms and ethical values. When specified as policy, guiding principles can enable AI stakeholders to form actionable, low-level requirements. Some requirements will be translated into quantitative measures of performance and effectiveness, while some may remain qualitative in nature.

Guiding principles that are relevant for AI risk include fairness, accountability, and transparency. Fairness in AI systems includes concerns for equality and equity by addressing socio-technical issues such as bias and discrimination. Individual human operators and their organizations should be answerable and held accountable for the outcomes of AI systems, particularly adverse

impacts stemming from risks. Absent transparency, users are left to guess about these factors and may make unwarranted and unreliable assumptions regarding model provenance. Transparency is often necessary for actionable redress related to incorrect and adverse AI system outputs.

5.3.1 Fairness

Standards of fairness can be complex and difficult to define because perceptions of fairness differ among cultures. For one type of fairness, process fairness, AI developers assume that ML algorithms are inherently fair because the same procedure applies regardless of user. However, this perception has eroded recently as awareness of biased algorithms and biased datasets has increased. Fairness is increasingly related to the existence of a harmful system, i.e., even if demographic parity and other fairness measures are satisfied, sometimes the harm of a system is in its existence. While there are many technical definitions for fairness, determinations of fairness are not generally just a technical exercise. Absence of harmful bias is a necessary condition for fairness.

5.3.2 Accountability

Determinations of accountability in the AI context are related to expectations for the responsible party in the event that a risky outcome is realized. Individual human operators and their organizations should be answerable and held accountable for the outcomes of AI systems, particularly adverse impacts stemming from risks. 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.

5.3.3 Transparency

Transparency seeks to remedy a common information imbalance between AI system operators and AI system consumers. Transparency reflects the extent to which information is available to a user when interacting with an AI system. Its scope spans from design decisions and training data to model training, the structure of the model, its intended use case, how and when deployment decisions were made and by whom, etc. Absent transparency, users are left to guess about these factors and may make unwarranted and unreliable assumptions regarding model provenance. Transparency is often necessary for actionable redress related to incorrect and adverse AI system outputs. A transparent system is not necessarily a fair, privacy-protective, secure, or robust system. However, it is difficult to determine whether an opaque system possesses such desiderata, and to do so over time as complex systems evolve.

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. The Core is composed of three elements: functions, categories, and subcategories. As illustrated in Figure 5, functions organize AI risk management activities at their highest level to map, measure, manage, and govern AI risks. Within each function, categories and subcategories subdivide the function into specific outcomes and actions.

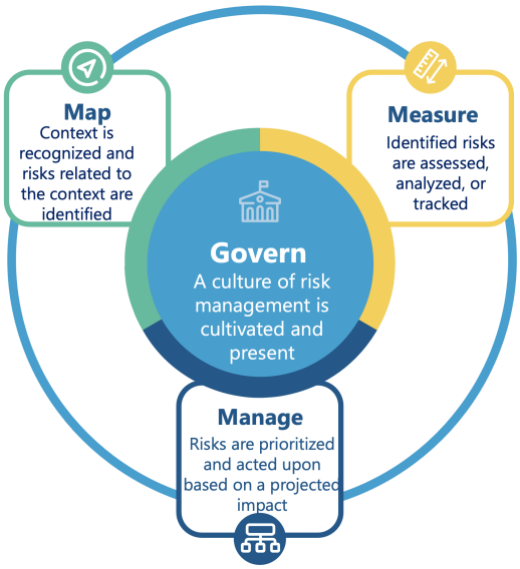

Figure 5: Functions organize AI risk management activities at their highest level to map, measure, manage, and govern AI risks. Governance is a cross-cutting function that is infused throughout and informs the other functions of the process.

Govern is a cross-cutting function that is infused throughout 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. 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. In most cases, it will be more useful and effective to begin with Map before Measure and Manage. Regardless, 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.

Technical and socio-technical characteristics and guiding principles of AI trustworthiness are essential considerations for each function. 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. Risk management should be performed throughout the AI system life cycle (Figure 6) to ensure it is continuous and timely.

On the following pages, Tables 1 through 4 provide the Framework Core listing.

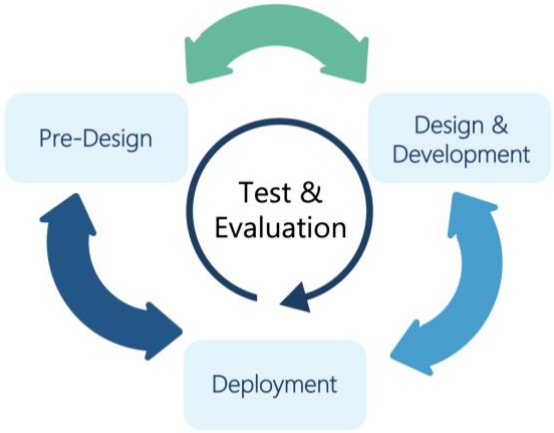

Figure 6: Risk management should be performed throughout the AI system life cycle to ensure it is continuous and timely. Example activities for each stage of the AI lifecycle follow. *Pre-Design*: data collection, curation or selection, problem formulation, and identification of stakeholders. *Design & Development*: data analysis, data cleaning, model training, and requirement analysis. *Test & Evaluation*: technical validation and verification. *Deployment*: user feedback and override, post deployment monitoring, and decommissioning.

6.1 Map

The Map function establishes the context and applies the attributes of the AI RMF taxonomy (Figure 3) to frame risks related to an AI system. The information gathered while carrying out this function informs decisions about model management, including 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 more formal quantitative analysis of benefits, costs, and risks.

A companion document describes practices related to mapping AI risks. Table 1 lists the Map function's categories and subcategories.

Table 1: Example of categories and subcategories for Map function

ID	Category	Subcategory
Map: Context is recognized and risks related to the context are identified		
1	Context is established and understood.	Intended purpose, setting in which the AI system will be deployed, the specific set of users along with their expectations, and impacts of system use are understood and documented as appropriate.
		The business purpose or context of use has been clearly defined or – in the case of assessing existing AI systems – re-evaluated.
		The organization's mission and relevant goals for the AI technology are understood.
		Stakeholders are defined, a plan for continuous engagement/communication is developed, and outreach is conducted.

		System requirements are elicited and understood from relevant stakeholders (e.g., "the system shall respect the privacy of its users"). Design decisions take socio-technical implications into account for addressing AI risks.
		Risk tolerances are determined.
2	Classification of AI system is performed.	The specific task that the AI system will support is defined (e.g., recommendation, classification, etc.).
		Considerations related to data collection and selection are identified. (e.g., availability, representativeness, suitability).
		Detailed information is provided about the operational context in which the AI system will be deployed (e.g., human-machine teaming, etc.) and how output will be utilized.
3	AI capabilities, targeted usage, goals, and expected benefits and costs over status quo are understood.	Benefits of intended system behavior are examined.
		Cost (monetary or otherwise) of errors or unintended system behavior is examined.
		Targeted application scope is specified and narrowed to the extent possible based on established context and AI system classification.
4	Risks and harms to individual, organizational, and societal perspectives are identified.	Potential business and societal (positive or adverse) impacts of technical and socio-technical characteristics for potential users, the organizations, or society as a whole are understood.
		Potential harms of the AI system are elucidated along technical and socio-technical characteristics and aligned with guiding principles.
		Likelihood of each harm is understood based on expected use, past uses of AI systems in similar contexts, public incident reports or other data.
		Benefits of the AI system outweigh the risks, and risks can be assessed and managed. Ideally, this evaluation should be conducted by an independent third party or by experts who did not serve as front-line developers for the system, and who consults experts, stakeholders, and impacted communities.

6.2 Measure

The Measure function provides knowledge relevant to the risks associated with attributes of the AI RMF taxonomy in Section 5. This includes analysis, quantitative or qualitative assessment, and tracking the risk and its impact. Risk analysis and measurement may involve a detailed consideration of uncertainties, tradeoffs, consequences, likelihood, events, controls, and their effectiveness. An event can have multiple causes and consequences and can affect multiple objectives.

Methods and metrics for quantitative or qualitative measurement rapidly evolve. Both qualitative and quantitative methods should be used to track risks.

A companion document describes practices related to measuring AI risks. Table 2 lists the Measure Function's categories and subcategories.

Table 2: Example of categories and subcategories for Measure function

ID	Category	Subcategory
Measure: Identified risks are assessed, analyzed, or tracked		
1	Appropriate methods and metrics are identified and applied.	Elicited system requirements are analyzed.
		Approaches and metrics for quantitative or qualitative measurement of the enumerated risks, including technical measures of performance for specific inferences, are identified and selected for implementation.
		The appropriateness of metrics and effectiveness of existing controls is regularly assessed and updated.
2	Systems are evaluated.	Accuracy, reliability, robustness, resilience (or ML security), explainability and interpretability, privacy, safety, bias, and other system performance or assurance criteria are measured, qualitatively or quantitatively.
		Mechanisms for tracking identified risks over time are in place, particularly if potential risks are difficult to assess using currently available measurement techniques, or are not yet available.
3	Feedback from appropriate experts and stakeholders is gathered and assessed.	Subject matter experts assist in measuring and validating whether the system is performing consistently with their intended use and as expected in the specific deployment setting.
		Measurable performance improvements (e.g., participatory methods) based on consultations are identified.

6.3 *Manage*

This function addresses risks which have been mapped and measured and are managed in order to maximize benefits and minimize adverse impacts. These are risks associated with the attributes of the AI RMF taxonomy (Section 5). Decisions about this function take into account the context and the actual and perceived consequences to external and internal stakeholders. That includes interactions of the AI system with the status quo world and potential benefits or costs.

Management can take the form of deploying the system as is if the risks are deemed tolerable; deploying the system in production environments subject to increased testing or other controls; or decommissioning the system entirely if the risks are deemed too significant and cannot be sufficiently addressed. Like other risk management efforts, AI risk management must be ongoing.

Practices related to AI risk management are discussed in the companion document. Table 3 lists the Manage function's categories and subcategories.

Table 3: Example categories and subcategories for Manage function

ID	Category	Subcategory
Manage: Risks are prioritized and acted upon based on a projected impact		
1	Assessments of potential harms and results of analyses conducted via the map and measure functions are used to respond to and manage AI risks.	Assessment of whether the AI is the right tool to solve the given problem (e.g., if the system should be further developed or deployed).
		Identified risks are prioritized based on their impact, likelihood, resources required to address them, and available methods to address them.

		Responses to enumerated risks are identified and planned. Responses can include mitigating, transferring or sharing, avoiding, or accepting AI risks.
2	Priority actions to maximize benefits and minimize harm are planned, prepared, implemented, and communicated to internal and external stakeholders as appropriate (or required) and to the extent practicable.	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.
		Plans are in place, both performance and control-related, to sustain the value of the AI system once deployed.
		Mechanisms are in place and maintained to supersede, disengage, or deactivate existing applications of AI that demonstrate performance or outcomes that are inconsistent with their intended use.
3	Responses to enumerated and measured risks are documented and monitored over time.	Plans related to post deployment monitoring of the systems are implemented, including mechanisms for user feedback, appeal and override, decommissioning, incident response, and change management.
		Measurable performance improvements (e.g., participatory methods) based on consultations are integrated into system updates.

6.4 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 processes focused on potential impacts of AI technologies are the backbone of risk management. Governance focuses on technical aspects of AI system design and development as well as on organizational practices and competencies that directly impact the individuals involved in training, deploying, and monitoring such systems. Governance should address supply chains, including third-party software or hardware systems and data as well internally developed AI systems.

Governance is a function that has relevance across all other functions, reflecting the importance of infusing governance considerations throughout risk management processes and procedures. Attention to governance is a continual and intrinsic requirement for effective AI risk management over an AI system's entire lifespan. For example, compliance with internal and external policies or regulations is a universal aspect of the governance function in risk management. Similarly, governance provides a structure through which AI risk management functions can align with organizational policies and strategic priorities, including those not directly related to AI systems.

A companion document describes practices related to governance of AI risk management. Table 4 lists Govern function's categories and subcategories.

Table 4: Example categories and subcategories for Govern function

ID	Category	Subcategory
Govern: A culture of risk management is cultivated and present		
1	Policies, processes, procedures and practices across the organization related to the development, testing, deployment, use and auditing of AI systems are in place, transparent, and implemented effectively.	The risk management process and its outcomes are documented and traceable through transparent mechanisms, as appropriate and to the extent practicable.
		Ongoing monitoring and periodic review of the risk management process and its outcomes are planned, with responsibilities clearly defined.
		Methods for ensuring all dimensions of trustworthy AI are embedded into policies, processes, and procedures.
2	Accountability structures are in place to ensure that the appropriate teams and individuals are empowered, responsible, and trained for managing the risks of AI systems.	Roles and responsibilities and lines of communication related to identifying and addressing AI risks are clear to individuals and teams throughout the organization.
		The organization's personnel and partners are provided AI risk management awareness education and training to enable them to perform their duties and responsibilities consistent with related policies, procedures, and agreements.
		Executive leadership of the organization considers decisions about AI system development and deployment ultimately to be their responsibility.
3	Workforce diversity, equity and inclusion processes are prioritized.	Decision making throughout the AI lifecycle is informed by a demographically and disciplinarily diverse team, including internal and external personnel. Specifically, teams that are directly engaged with identifying design considerations and risks include a diversity of experience, expertise and backgrounds to ensure AI systems meet requirements beyond a narrow subset of users.
4	Teams are committed to a culture that considers and communicates risk.	Teams are encouraged to consider and document the impacts of the technology they design and to develop and communicate about these impacts more broadly.
		Organizational practices are in place to ensure that teams actively challenge and question steps in the design and development of AI systems to minimize harmful impacts.
5	Processes are in place to ensure that diversity, equity, inclusion, accessibility, and cultural considerations from potentially impacted individuals and communities are fully taken into account.	Organizational policies and practices are in place that prioritize the consideration and adjudication of external stakeholder feedback regarding the potential individual and societal harms posed by AI system deployment.
		Processes are in place to empower teams to make decisions about if and how to develop and deploy AI systems based on these considerations, and define periodic reviews of impacts, including potential harm.
6	Clear policies and procedures are in place to address AI risks arising from supply chain issues, including third-party software and data.	Policies and procedures include guidelines for ensuring supply chain and partner involvement and expectations regarding the value and trustworthiness of third-party data or AI systems.
		Contingency processes are in place to address potential issues with third-party data or AI systems.

7 AI RMF Profiles

Profiles are instantiations of the AI RMF Core for managing AI risks for context-specific use cases. Using the AI RMF, profiles illustrate how risk can be managed at various stages of the AI lifecycle or in sector, technology, or end-use applications. Profiles may state an "as is" and "target" state of how an organization addresses AI risk management.

NOTE: Development of profiles is deferred until later drafts of the AI RMF are developed with the community. NIST welcomes contributions of AI RMF profiles. These profiles will inform NIST and the broader community about the usefulness of the AI RMF and likely lead to improvements which can be incorporated into future versions of the framework.

8 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 in order to maximize benefits and to minimize AI-related harms. Organizations 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.

NOTE: NIST is deferring development of this section until later drafts of the AI RMF are developed with the community.

Part 3: Practice Guide

9 Practice Guide

NOTE: NIST is developing a companion Practice Guide which will include additional examples and practices that can assist in using the AI RMF. That Guide, which will reside online only and will be updated regularly with contributions expected to come from many stakeholders, will be part of the NIST AI Resource Center that is being established.