Understanding NIST’s AI Risk Management Framework

April 26, 2023

Ben Rossen
Special Counsel

BAKER BOTTS

Table of Contents

Overview of Artificial Intelligence

- What is AI?
- Relevant Regulatory and Legislative Developments

Introduction to the NIST AI RMF 1.0

- Background on the NIST Framework
- Goals, limitations, and future plans for the Framework

Framing Risks, AI Harms, and Trustworthy AI Systems

- Risks and harms that AI systems should address
- Characteristics of trustworthy AI systems

How to Manage Risk with AI RMF 1.0 and the Playbook

- Key principles from the Framework
- How the Framework aligns with regulator expectations and guidance

OVERVIEW OF ARTIFICIAL INTELLIGENCE TECHNOLOGIES AND REGULATION

What is AI?

- In the 1950s, Alan Turing posed the question “Can machines think?” and developed the “Turing Test.”
  - AI has evolved quite a bit since then, but the fundamental technologies are not new
- AI combines computer science and large datasets to enable problem-solving
  - Encompasses numerous subfields:
    - Machine learning, deep learning, computer vision, neural networks
- “Weak AI” vs. “Strong AI”
  - Weak AI: trained and focused to perform specific, narrow tasks
    - Siri, Alexa, current AVs
  - Strong AI: Artificial General Intelligence (AGI) & Artificial Super Intelligence (ASI)
    - AGI: theoretical AI where machine would have an intelligence equal to humans
    - ASI: AI that surpasses the intelligence and ability of the human brain

AI Timeline

Key dates and developments

1950

Alan Turing publishes *Computing Machinery and Intelligence*

1956

John McCarthy coins the term “artificial intelligence” at the first ever AI conference

1967

First computer trained on a neural network that “learned” through trial and error

1997

IBM’s Deep Blue beats Gary Kasparaov in a chess match

2011

IBM Watson beats Ken Jennings at Jeopardy!

2016

DeepMind’s AlphaGo program defeats world champion Go player

2018

OpenAI publishes research paper on generative pre-training (GPT) of a language model

2019

OpenAI releases GPT-2 language model

2022

Stability AI develops Stable Diffusion, a deep learning text-to-image model, leading to other image services like DALL-E and Midjourney

2022

ChatGPT released, reaches one million users within five days

2023

GPT-4 released, Microsoft integrates ChatGPT into Bing, Google releases Bard, generative AI explodes…

AI Terminology

Algorithms, Models, Deep Learning, AI Systems

- **Algorithm** – the AI algorithm is a procedure that runs on a dataset to recognize patterns, rules, etc.
- **Model** – Models are essentially the output of the algorithm once it is run through the data (often millions of times)
  - The effectiveness of the algorithm's training will determine the precision and confidence of the model
- **Machine Learning** – ML uses algorithms to parse data, learn from it, and make decisions based on what it has learned
- **Deep Learning** – A subset of ML that uses neural networks to mimic the learning process of the human brain
  - Deep learning technology drives much of the advances in AI: image and speech recognition, natural language processing, e.g.
- **AI Systems**
  - NIST's definition: An engineered or machine-based system that can, for a given set of objectives, generate outputs such as predictions, recommendations, or decisions influencing real or virtual environments, designed to operate with varying levels of autonomy.

Regulatory & Self-Regulatory Context

- “Hard law” vs. “Soft law”
  - Hard law = enforceable laws created by government (statutes, regs, treaties, etc.)
  - Soft law = substantive requirements not directly enforceable
    - Principles, guidelines, standards, codes of conduct, voluntary programs, etc.

- US laws applicable to AI
  - Section 5 of the FTC Act (UDAP)
  - FCRA, ECOA, Civil rights law
  - State comprehensive privacy laws (e.g., California rulemaking)
  - State and Local laws (employment, facial recognition)

- EU AI Act (pending)
- GDPR

02

INTRO TO THE NIST AI RMF 1.0

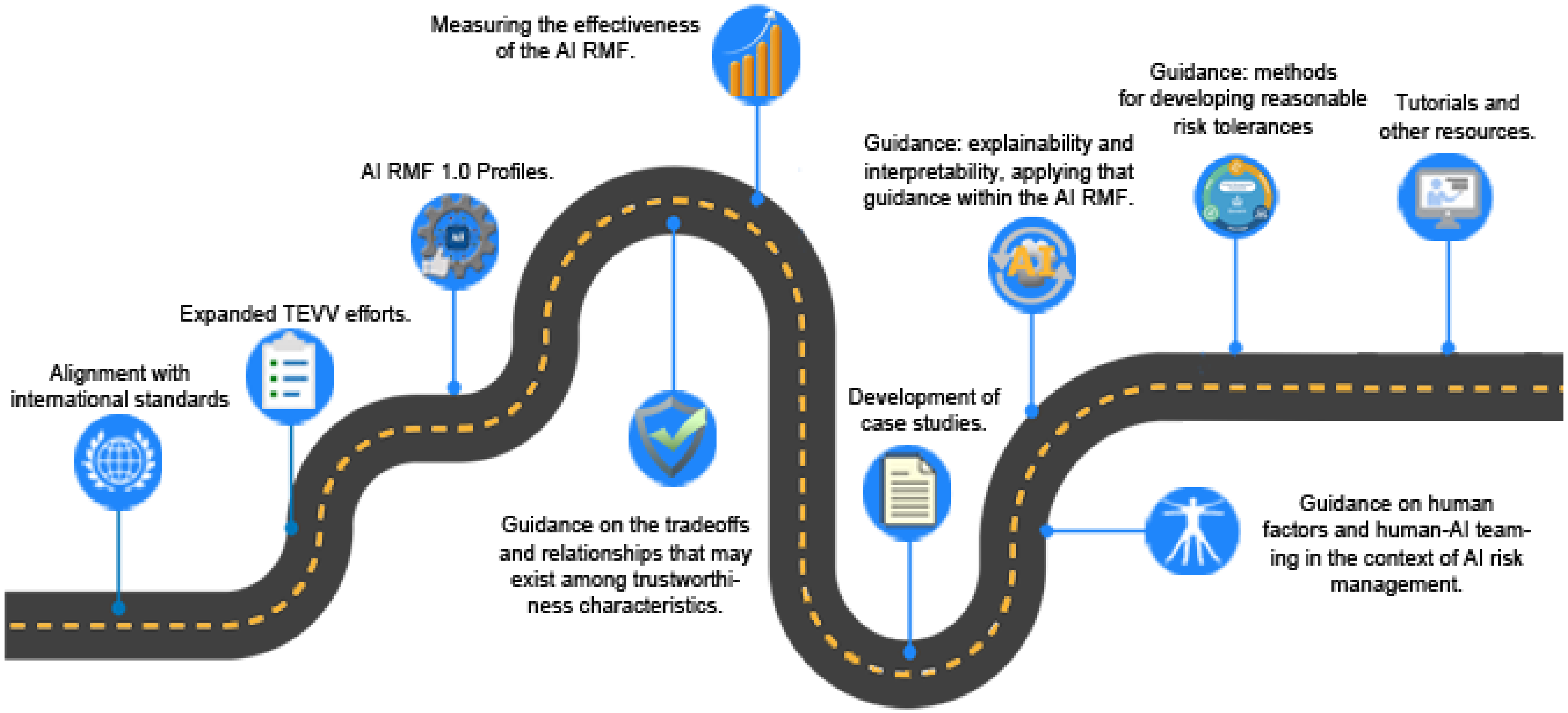
Background

- National Artificial Intelligence Initiative Act of 2020 (P.L. 116-283)
- A **voluntary** framework that can be applied by organizations of all sizes and in all sectors to implement approaches to AI risk management
- Characteristics of the Framework
  - A **practical** approach
  - Intended to **adapt over time** as technologies develop
- A living document
  - NIST intends to update the Framework and supporting resources based on evolving technology, the global standards landscape, and AI community experience and feedback
  - AI RMF is intended to align with international standards, guidelines, and practices

Components of the AI RMF 1.0

- AI Risk Management Framework
  - Framing Risk
    - Understanding and addressing risks, impacts, and harms
    - Challenges for AI Risk Management
  - Audience
    - Understanding the AI lifecycle and the relevant AI actors at all stages of development, deployment, and use
  - AI Risks and Trustworthiness
    - Unique risks of AI systems
    - Understanding the characteristics of trustworthy AI systems
  - Effectiveness
    - The expected benefits for users of the AI RMF 1.0
  - AI RMF Core
  - AI RMF Profiles
- AI RMF Playbook

Roadmap



Alignment with international standards

Expanded TEVV efforts.

AI RMF 1.0 Profiles.

Measuring the effectiveness of the AI RMF.

Guidance on the tradeoffs and relationships that may exist among trustworthiness characteristics.

Development of case studies.

Guidance: explainability and interpretability, applying that guidance within the AI RMF.

Guidance: methods for developing reasonable risk tolerances

Tutorials and other resources.

Guidance on human factors and human-AI teaming in the context of AI risk management.

Goals of the RMF

- Not a checklist – it's a set of processes to help organizations think about risk
- Framework users may benefit from:
  - Enhanced processes for governing and managing risk, documenting outcomes
  - Improved awareness of the relationships and trade-offs of trustworthiness characteristics
  - Explicit processes for determining go/no-go for deployment decisions
  - Organizational accountability
  - Developing a culture of risk and impact assessment
  - Improved information sharing within and across organizations
  - Greater knowledge of downstream risk
  - Better engagement with AI actors and interested parties
  - Better capacity for testing, evaluation, validating and verifying AI risk

03

FRAMING RISKS, AI HARMS, AND TRUSTWORTHY AI SYSTEMS

Risk

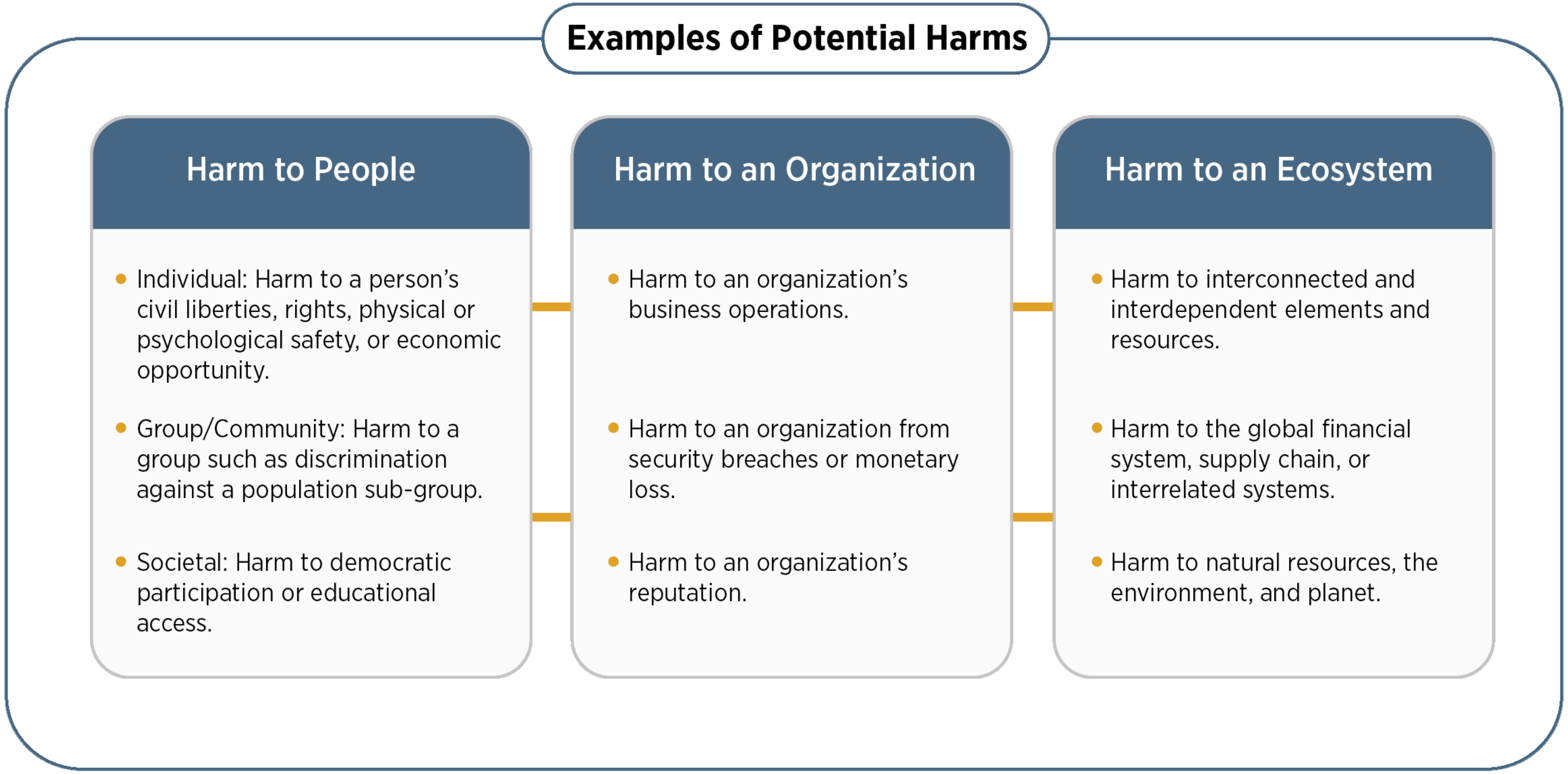
- **Risk** refers to the composite measure of event’s probability of occurring and the magnitude or degree of the consequences of the corresponding event.
  - When evaluating negative impact, risk is a function of:
    - The negative impact or magnitude of harm if an event occurs; and
    - The likelihood of occurrence

(Adapted from OMB Circular A-130:2016)

- **Risk management** refers to coordinated activities to direct and control an organization with regard to risk

(Adapted from ISO 310000:2018)

Potential AI Harms



Examples of Potential Harms

Harm to People

- Individual: Harm to a person’s civil liberties, rights, physical or psychological safety, or economic opportunity.
- 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

- 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 an 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.

Difficulties in Assessing AI Risk

Third-party software, hardware, and data

- Developers may evaluate risk differently from the organization deploying or operating the system, they may lack transparency, and customers may use or integrate 3rd party systems without appropriate safeguards

Emergent Risks

- Organizations need a system to track risks as they emerge and techniques for measuring them

Reliable Metrics

- There is a lack of consensus on verifiable measurement methods for risk and trustworthiness
- Measurement may lack nuance, fail to account for sub-groups and context of use

Risks at Different Stages of the AI Lifecycle

- Risks can vary at different stages of development
- E.g., a developer of pre-trained models can have a different risk profile than the person deploying that model in a specific use case

Difficulties in Assessing AI Risk

Risks in real-world settings

- Measuring risks in a laboratory or controlled environment may differ from risks that emerge in operational settings

Inscrutability

- Opaque systems with limited explainability or interpretability pose challenges
- AI systems may lack transparency or documentation
- Inherent uncertainty of AI systems

Human baseline

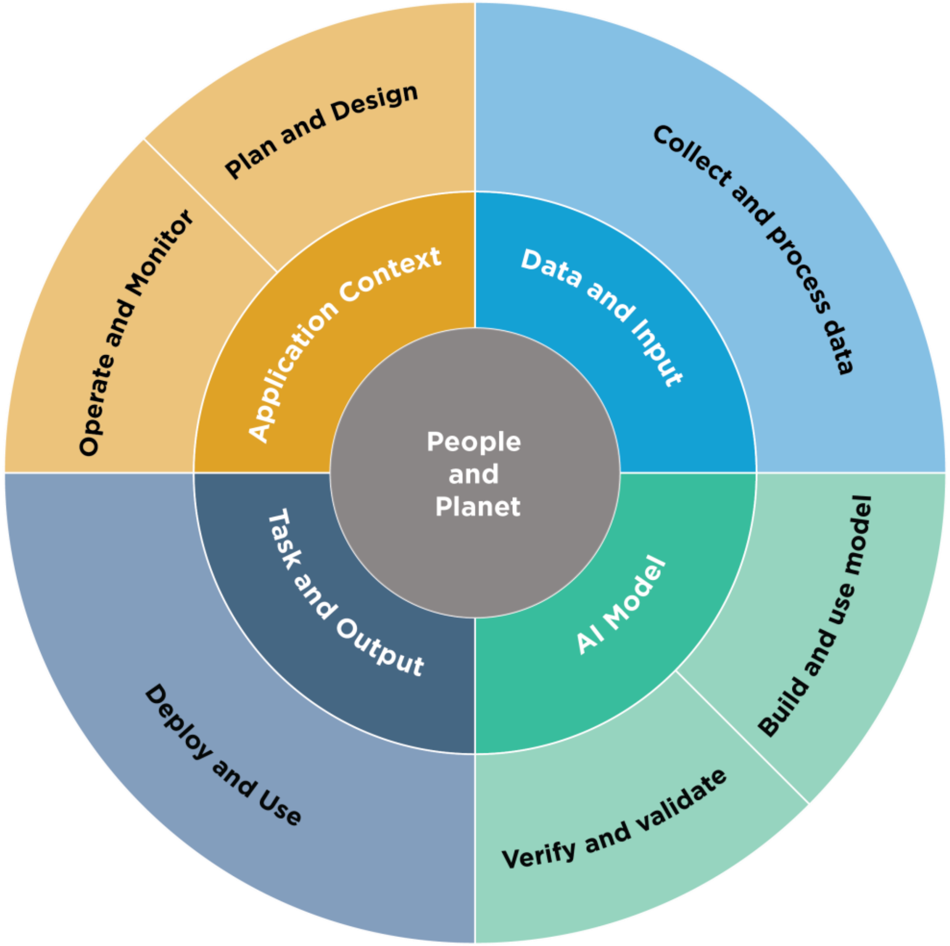
- AI systems intended to replace or augment human activity or decision-making need baseline metrics for comparison

Risk Prioritization vs. Risk Tolerance

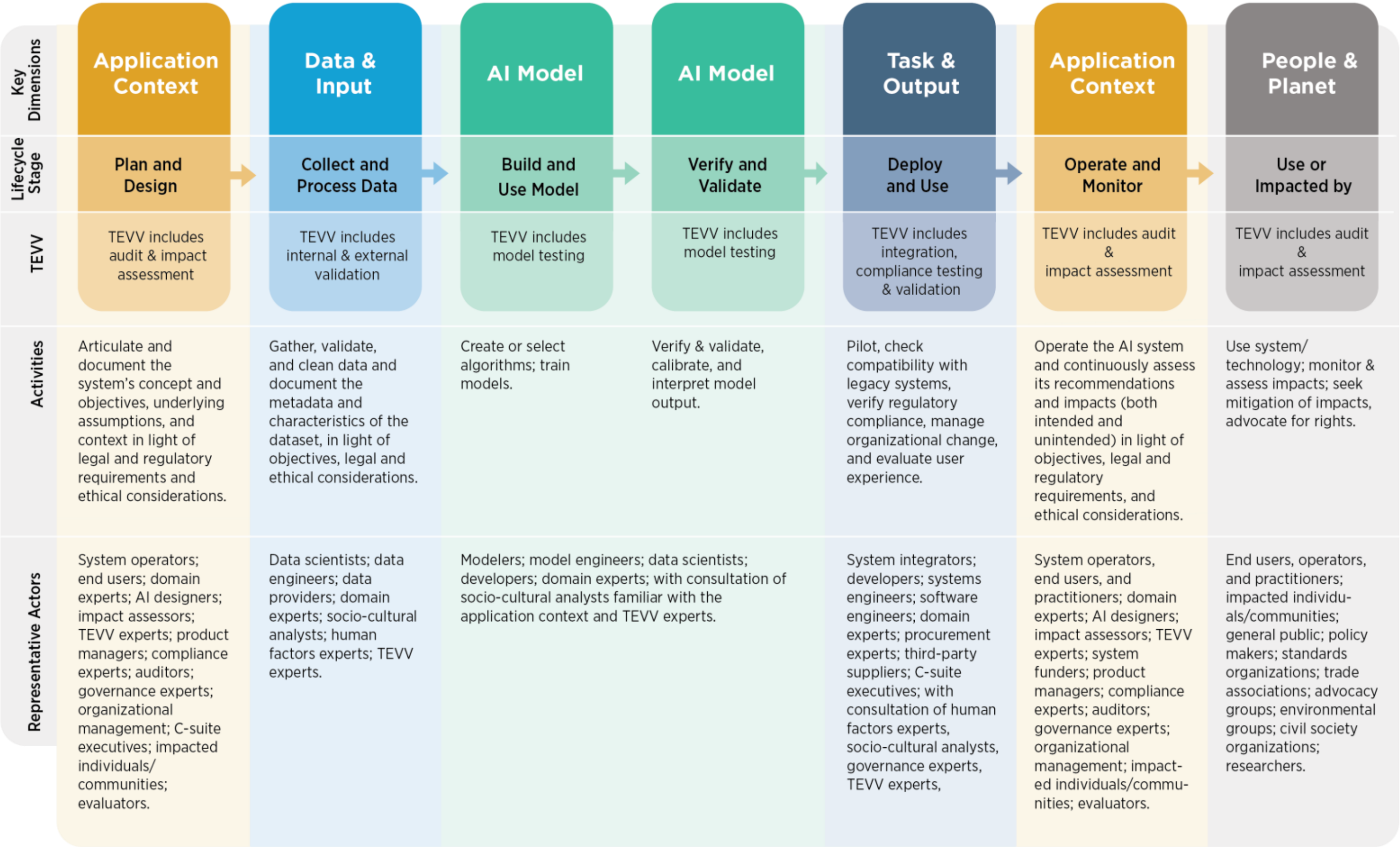
- The AI RMF helps organizations *prioritize* risk management
- ***It does not prescribe risk tolerance***
  - Acceptable risk tolerance is highly contextual and specific to particular applications and uses
  - Risk tolerance depends on legal or regulatory requirements, and can change over time as AI systems, policies, and norms evolve
- ***It is frequently impossible to eliminate all risk***
  - A risk management culture helps organizations understand that not all risks are the same, that resources must be allocated purposefully and prioritized based on the level of risk and the potential impact of an AI system
  - Residual risks (the risk remaining after risk treatment) directly impact end users, individuals, and communities
    - Understanding residual risk will help AI system providers consider whether a system should be deployed

Audience

- This image depicts the AI lifecycle and the primary audience for each stage of the testing, evaluation, verification, and validation (TEVV)
- The inner circle identifies the AI actors who perform or manage the design, development, deployment, evaluation, and use of AI systems – the *primary audience* of AI RMF
- People & Planet at the center represent human rights and broader well-being of society and the planet, who inform the primary audience
    - Advocacy groups, civil society, researchers, standards developers, etc.

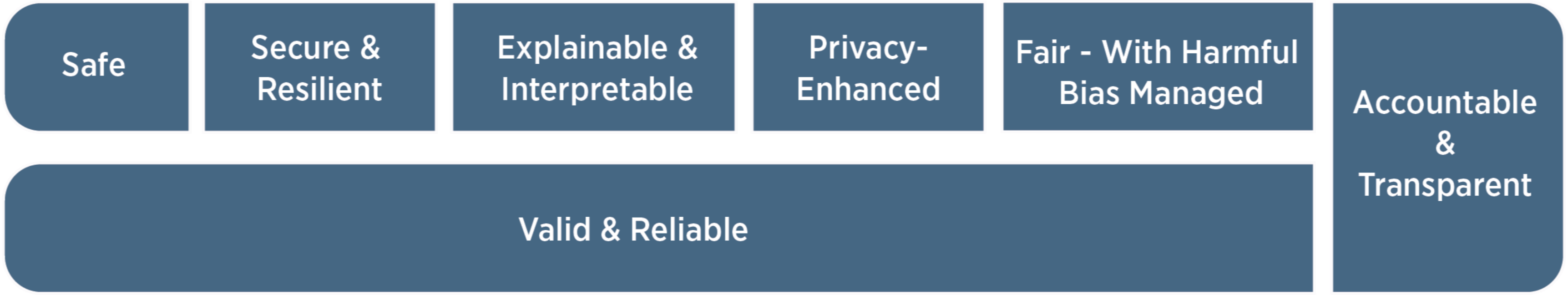


AI Actors and the AI Development Lifecycle



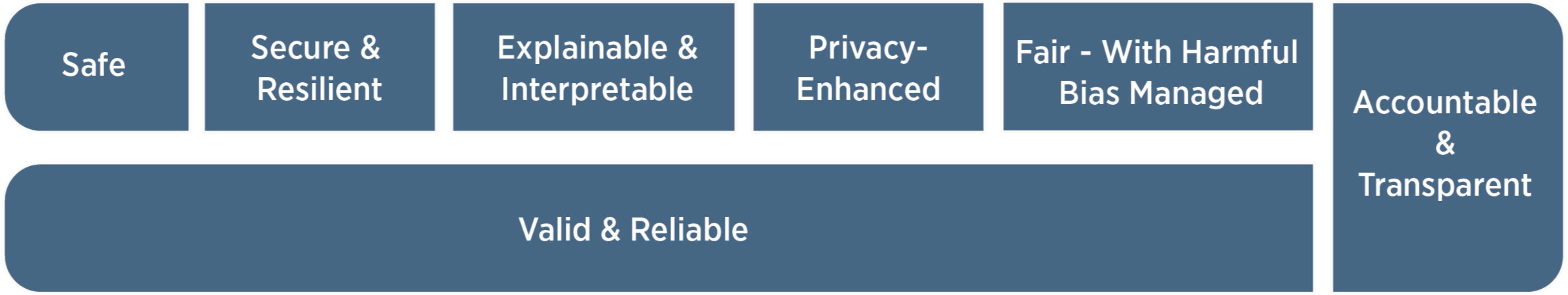
Key Dimensions	Application Context	Data & Input	AI Model	AI Model	Task & Output	Application Context	People & Planet
Lifecycle Stage	Plan and Design	Collect and Process Data	Build and Use Model	Verify and Validate	Deploy and Use	Operate and Monitor	Use or Impacted by
TEVV	TEVV includes audit & impact assessment	TEVV includes internal & external validation	TEVV includes model testing	TEVV includes model testing	TEVV includes integration, compliance testing & validation	TEVV includes audit & impact assessment	TEVV includes audit & impact assessment
Activities	Articulate and document the system's concept and objectives, underlying assumptions, and context in light of legal and regulatory requirements and ethical considerations.	Gather, validate, and clean data and document the metadata and characteristics of the dataset, in light of objectives, legal and ethical considerations.	Create or select algorithms; train models.	Verify & validate, calibrate, and interpret model output.	Pilot, check compatibility with legacy systems, verify regulatory compliance, manage organizational change, and evaluate user experience.	Operate the AI system and continuously assess its recommendations and impacts (both intended and unintended) in light of objectives, legal and regulatory requirements, and ethical considerations.	Use system/ technology; monitor & assess impacts; seek mitigation of impacts, advocate for rights.
Representative Actors	System operators; end users; domain experts; AI designers; impact assessors; TEVV experts; product managers; compliance experts; auditors; governance experts; organizational management; C-suite executives; impacted individuals/ communities; evaluators.	Data scientists; data engineers; data providers; domain experts; socio-cultural analysts; human factors experts; TEVV experts.	Modelers; model engineers; data scientists; developers; domain experts; with consultation of socio-cultural analysts familiar with the application context and TEVV experts.		System integrators; developers; systems engineers; software engineers; domain experts; procurement experts; third-party suppliers; C-suite executives; with consultation of human factors experts, socio-cultural analysts, governance experts, TEVV experts,	System operators, end users, and practitioners; domain experts; AI designers; impact assessors; TEVV experts; system funders; product managers; compliance experts; auditors; governance experts; organizational management; impacted individuals/communities; evaluators.	End users, operators, and practitioners; impacted individuals/communities; general public; policy makers; standards organizations; trade associations; advocacy groups; environmental groups; civil society organizations; researchers.

Characteristics of Trustworthy AI Systems


Safe
Secure & Resilient
Explainable & Interpretable
Privacy-Enhanced
Fair - With Harmful Bias Managed
Valid & Reliable
Accountable & Transparent

- Creating trustworthy AI systems requires balancing each of these characteristics based on the AI system's context of use
- Balancing usually requires trade-offs
  - All characteristics rarely apply in every setting
  - Some can be more or less important than others depending on context and use

Characteristics of Trustworthy AI Systems

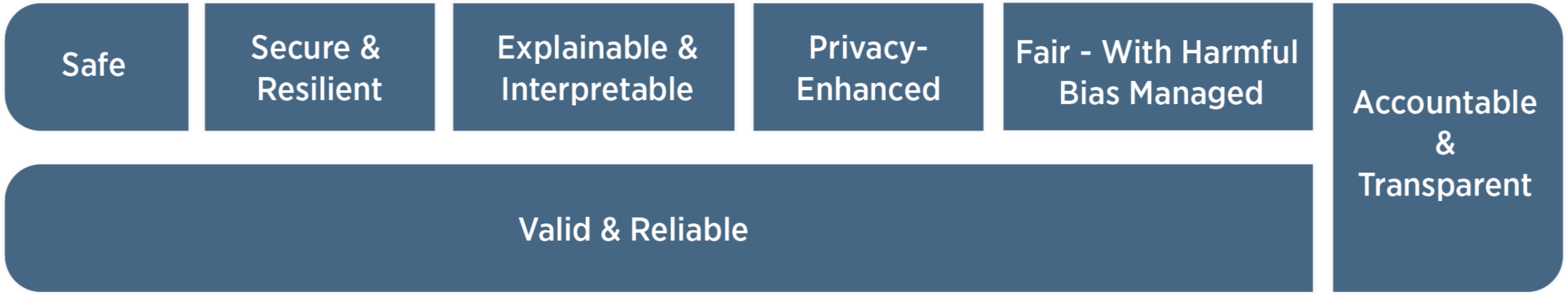

Safe
Secure & Resilient
Explainable & Interpretable
Privacy-Enhanced
Fair - With Harmful Bias Managed
Accountable & Transparent
Valid & Reliable

- Examples of undesirable systems
  - Highly secure but unfair
  - Accurate but opaque and uninterpretable
  - Inaccurate but secure, privacy-enhanced and transparent
- 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 interested parties

Valid and Reliable

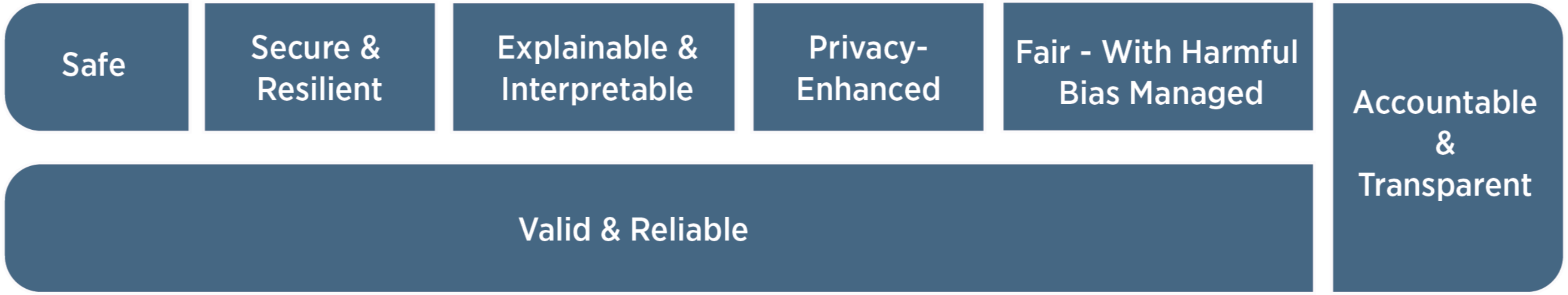
- *Validation* is the “confirmation, through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled” (ISO 9000:2015)
- *Reliability* is the “ability of an item to perform as required, without failure, for a given time interval, under given conditions” (ISO/IEC TS 5723:2022)
- *Accuracy* and *Robustness* contribute to validity and can be in tension in AI systems
  - Accuracy is the “closeness of results of observations, computations, or estimates to the true values or the values accepted as being true.”
  - Robustness is the “ability of a system to maintain its level of performance under a variety of circumstances”
    - This means a system performs correctly in expected uses, and minimizes potential harms if it is operating in an unexpected setting
- Valid and Reliable systems are assessed by ongoing testing or monitoring that confirms a system is working as intended

Safe


Safe
Secure & Resilient
Explainable & Interpretable
Privacy-Enhanced
Fair - With Harmful Bias Managed
Accountable & Transparent
Valid & Reliable

- AI systems should "not under defined conditions, lead to a state in which human life, health, property, or the environment is endangered"
  (ISO/IEC TS 5723:2022)
- Safe operation is improved through:
  - Responsible design, development, and deployment
  - Clear information to deployers about responsible use
  - Responsible decision-making by deployers and end-users
  - Documentation of risks based on empirical evidence of incidents
- Simulations, in-domain testing, real-time monitoring, ability to shut down, modify, or have human intervention

Secure & Resilient



Safe

Secure & Resilient

Explainable & Interpretable

Privacy-Enhanced

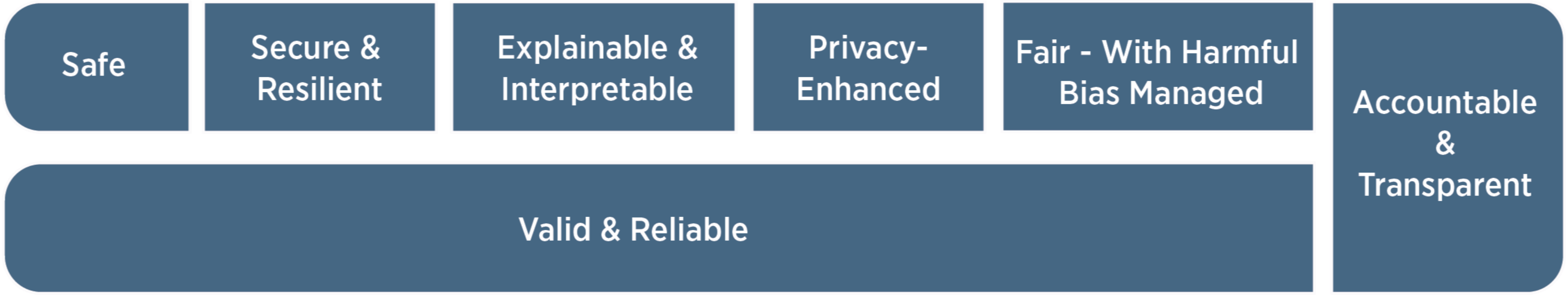
Fair - With Harmful Bias Managed

Accountable & Transparent

Valid & Reliable

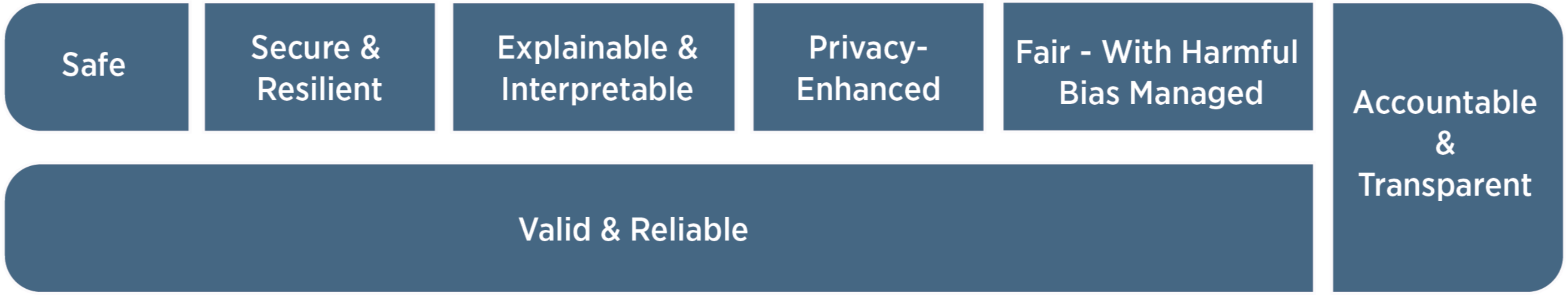
- *Resilient* systems can withstand unexpected adverse events or unexpected changes in their environment or use
- *Secure* systems maintain confidentiality, integrity, and availability through mechanisms that prevent unauthorized access

Accountable and Transparent


Safe
Secure & Resilient
Explainable & Interpretable
Privacy-Enhanced
Fair - With Harmful Bias Managed
Valid & Reliable
Accountable & Transparent

- *Accountable* systems presuppose transparency
- *Transparency* reflects the extent to which information about an AI system and its outputs is available to individuals interacting with the system
  - It covers the entire lifecycle from design decisions, training data, model training, model structure, intended use cases, etc.
  - Transparent systems are not necessarily accurate, secure, or fair, but it is difficult or impossible to determine whether an opaque system possesses such characteristics
- NIST encourages developers of AI systems to test different types of transparency tools for AI deployers to ensure systems are used as intended

Explainable & Interpretable



Safe

Secure & Resilient

Explainable & Interpretable

Privacy-Enhanced

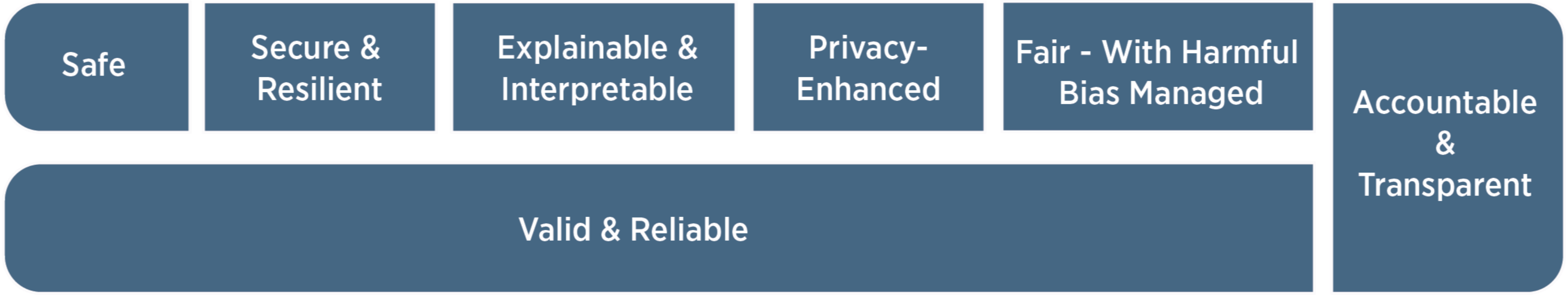
Fair - With Harmful Bias Managed

Accountable & Transparent

Valid & Reliable

- *Explainability* refers to the mechanisms underlying the operation of an AI system ("how" a decision was made)
  - Explainable systems can be debugged and monitored more easily
  - They facilitate thorough documentation, audit, and governance
- *Interpretability* refers to the meaning of an AI system's output in the context of its designed functional purpose ("why" a decision was made)
  - Interpretability risks can often be addressed by describing why an AI system made a particular prediction or recommendation

Privacy-Enhanced



Safe

Secure & Resilient

Explainable & Interpretable

Privacy-Enhanced

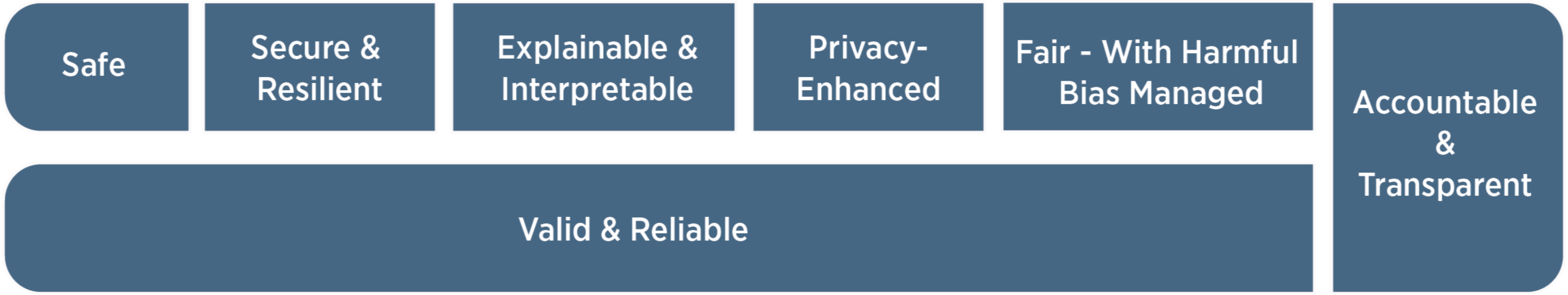
Fair - With Harmful Bias Managed

Accountable & Transparent

Valid & Reliable

- Norms and practices that help to safeguard human autonomy, identity and dignity
  - Privacy values such as anonymity, confidentiality, and control can guide AI system design
  - AI systems present new privacy risks by allowing inference to identify individuals or previously private information about individuals
- PETs, data minimization for certain model outputs (de-identification/aggregation) can support privacy, but often at a cost in accuracy

Fair – With Harmful Bias Managed


Safe
Secure & Resilient
Explainable & Interpretable
Privacy-Enhanced
Fair - With Harmful Bias Managed
Accountable & Transparent
Valid & Reliable

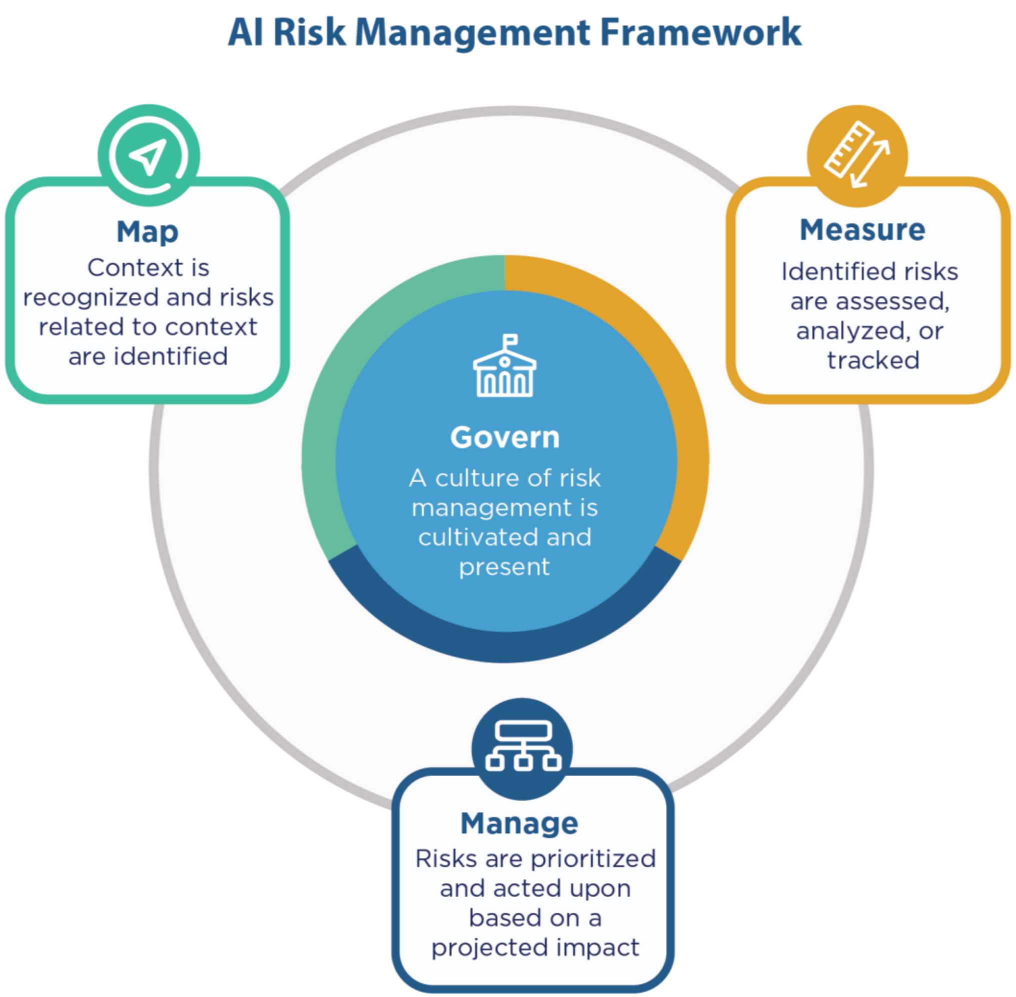
- *Fairness* includes concerns for equality and equity
  - Harmful bias and discrimination
- *Bias* means more than just demographic balance and data representativeness
  - *Systemic bias*
  - *Computational/statistical bias*
  - *Human-cognitive bias*
- AI systems can potentially increase the speed and scale of biases, perpetuating and amplifying harms to individuals, groups, communities, etc.

04

MANAGING AI RISKS WITH THE RMF 1.0 CORE AND THE PLAYBOOK

AI RMF Core

- Govern
- Map
- Measure
- Manage



AI Risk Management Framework

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

GOVERN

- The GOVERN function cuts across the AI lifecycle and enables all the other functions of RMF
- The GOVERN function
  - Cultivates a culture of risk management
  - Outlines processes and procedures to identify and manage risk
  - Incorporates processes to assess impacts
  - Provides a structure for risk management functions to align with organizational principles
  - Addresses full product lifecycle

GOVERN (cont’d)

- The GOVERN function includes:
  - Policies and procedures to map, measure, and manage risk
  - Accountability structures for teams and individuals, with clearly defined roles, appropriate training, and responsible senior leadership
  - Workforce diversity and inclusion with a variety of experience, demographics, disciplines, and backgrounds
  - Policies and procedures that facilitate a culture that considers and communicates AI risk
  - Policies for engagement with AI actors and to incorporate feedback into system design and implementation
  - Policies and procedures to address AI risk from third-parties software and data

MAP

- The MAP function enables all actors throughout the AI lifecycle to have the context necessary to frame risk
  - Decisions made at one stage of the lifecycle can be undermined by decisions made in later stages by other AI actors, introducing uncertainty
- Mapping enables negative risk prevention and can inform decision-making about the development and use of models and whether an AI solution is appropriate or needed
- The MAP function assists with:
  - Improving capacity for understanding contexts of an AI system
  - Checking assumptions about context and use
  - Enabling recognition of when systems do not function within or outside of their expected context or use
  - Understanding limitations
  - Identifying constraints of real-world applications that can lead to negative risk
  - Identifying known and foreseeable impacts
  - Anticipating risks beyond an AI system’s intended use

MAP (cont’d)

- The MAP function includes:
  - Establishing and understanding the context of an AI system and its requirements
    - Intended purposes, beneficial uses, positive and negative impacts on individuals/society, assumptions and limitations, etc.
  - Categorization of an AI system
    - Define the specific tasks and methods to implement the AI system
  - Understanding AI capabilities, usage, goals, costs/benefits
  - Mapping risks and benefits for all components of the system and data, including third-party software
  - Characterizing impacts to individuals, groups, communities, and society

MEASURE

- The MEASURE function uses quantitative, qualitative, or mixed-method metrics to analyze, assess, benchmark, and monitor AI risk and related impact
- It helps organizations identify trade-offs among trustworthiness characteristics and consider options to address them
- Characteristics of the MEASURE function:
  - Objective, repeatable, scalable TEVV processes
  - Methodologies adhering to legal, scientific, and ethical norms
  - Open and transparent processes
- Measurement outcomes are used by the MANAGE function to assist monitoring and response efforts

MEASURE (cont’d)

- The MEASURE function includes:
  - Identifying appropriate metrics and methodologies
  - Documenting risks that cannot or will not be measured
  - Evaluating AI systems for trustworthy characteristics, such as
    - Regular checks for safety concerns
    - Identifying AI system security and resilience
    - Explaining and validating AI models and output
    - Explaining privacy risks of AI system
    - Evaluating fairness/bias
  - Processes for tracking risks over time and incorporating feedback

MANAGE

- The MANAGE function involves allocating resources to address risks that have been MAPPED and MEASURED on a regular basis, using processes defined in the GOVERN function
- It includes:
  - Prioritizing, responding to, and addressing AI risks based on assessments and analytical output from the MAP and MEASURE functions
  - Implementing strategies to maximize AI benefits and minimize negative impacts based on input from all relevant AI actors
  - Managing risks and benefits of third-party software and entities
  - Documenting and monitoring risk treatments, response and recovery plans
  - Monitoring AI systems post-deployment and implementing activities for continuous improvement

The RMF Playbook

- The AI RMF Playbook helps organizations implement the CORE functions
- Downloadable in PDF, Excel/CSV, or JSON
- Highly filterable based on the different CORE functions, the AI actors involved and specific topics
- Each section of the playbook gives more detail about each subsection of every CORE function
  - Provides a list of potential suggested actions to implement
  - Identifies issues that organizations may need to document and ask themselves
  - Provides extensive additional resources drawn from regulators, guidance, ethics frameworks, research, and AI principles
- Will be regularly updated by NIST going forward

Questions?

AUSTIN
BRUSSELS
DALLAS
DUBAI
HOUSTON
LONDON
NEW YORK
PALO ALTO
RIYADH
SAN FRANCISCO
SINGAPORE
WASHINGTON

bakerbotts.com