datacamp

NIST AI Risk Management Framework

(2026)

What Is It?

NIST AI RMF is a framework for identifying and managing AI risks across the full system lifecycle. It is applicable to any organization in any industry. The framework is intended to be continuously applied, not just a one-time checklist.

Harm to People

- Individual: civil liberties, rights, physical/ psychological safety, economic opportunity.
- Group/Community: discrimination against a population sub-group.
- Societal: democratic participation or educational access.

Harm to an Organization

- Harm to an organization's business operations.
- Harm from security breaches or monetary loss.
- Harm to an organization's

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.

Fig. 1 — Potential harms related to AI systems

What Makes AI Trustworthy?

All seven must be balanced based on context. They sometimes conflict and tradeoffs must be managed, not ignored. Trustworthiness is only as strong as its weakest characteristic.

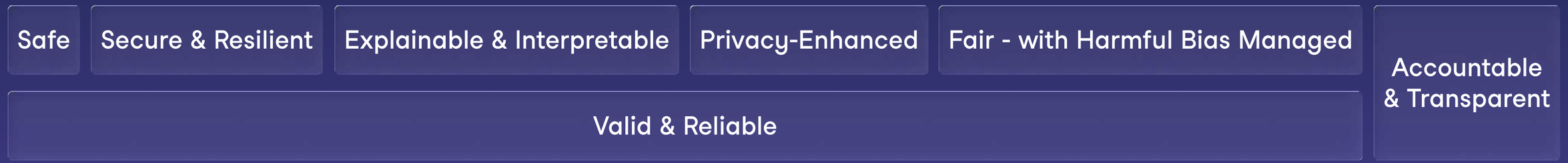


Fig. 2 — Trustworthy AI characteristics hierarchy

What Makes AI Trustworthy?

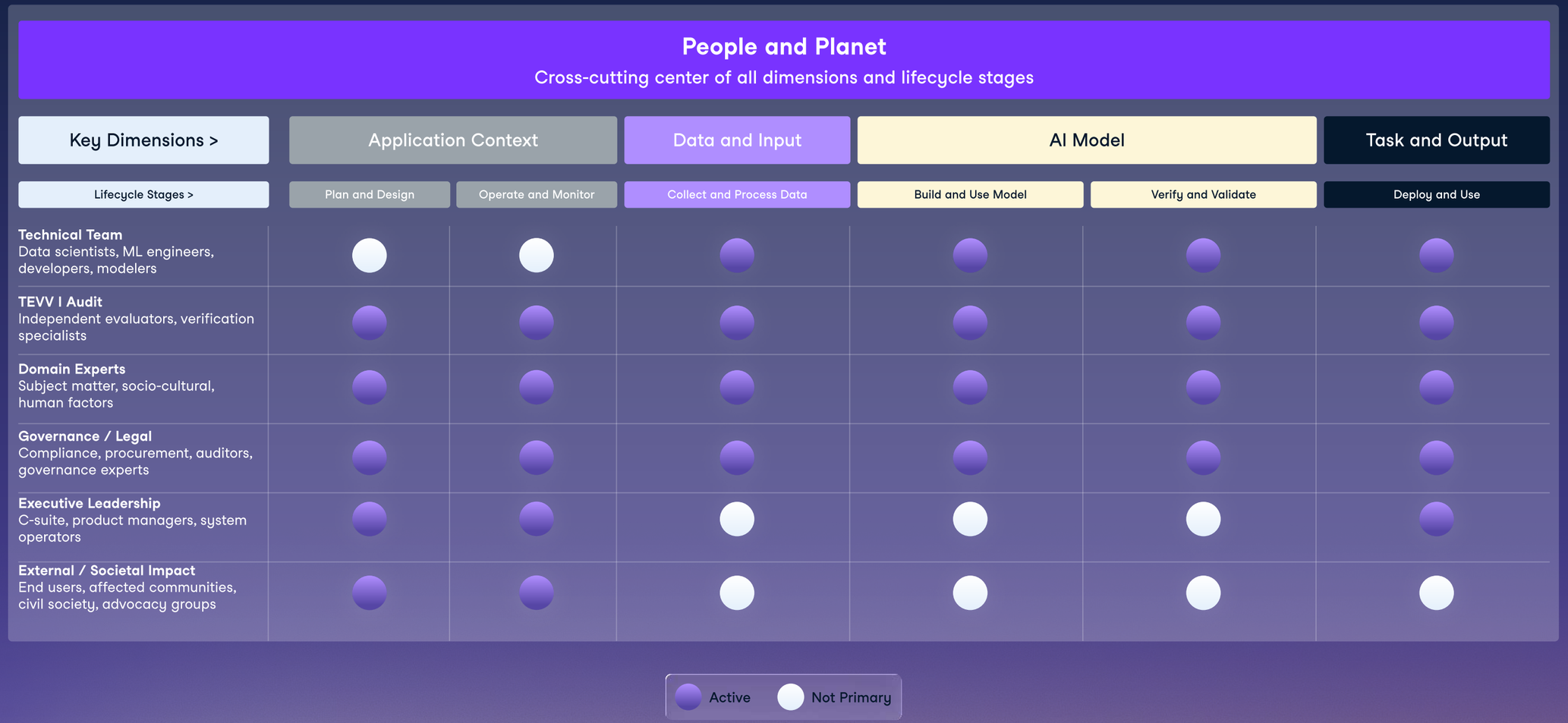
Characteristic	Description
Valid & Reliable	The necessary foundation. Performs consistently across real-world conditions, not just benchmarks. Covers accuracy, robustness, and generalizability.
Accountable & Transparent	Cross-cutting across all characteristics. Clear audit trails. Users know they are interacting with AI. Decision logs accessible for review.
Safe	Does not cause physical, psychological, or societal harm. Includes fail-safes and human override mechanisms.
Secure & Resilient	Withstands adversarial attacks, prompt injection, and data poisoning. Degrades safely when needed.
Explainable & Interpretable	Explainability covers how a decision was made. Interpretability covers why it matters. Critical for legal, financial, and HR contexts.
Privacy-Enhanced	Compliant with GDPR, CCPA, and equivalents. Applies data minimization and privacy-preserving techniques.
Fair: With Harmful Bias Managed	Actively identifies and reduces discriminatory outcomes across three bias categories: systemic, computational/statistical, and human-cognitive.

Terminology

Term	Definition
Socio-technical	Risk is not just about code. It emerges from how society uses the system, who operates it, and the deployment context.
Risk Tolerance	An organization's readiness to bear risk to achieve its objectives. Highly contextual and use-case specific.
Residual Risk	Risk remaining to end users and affected communities after all controls have been applied.
TEVV	Test, Evaluation, Verification & Validation. How you measure that your AI is trustworthy.
Context of Use	The specific conditions, users, and purposes of deployment. The same model carries different risk profiles in different deployments.

AI system lifecycle

AI systems have six stages in their lifecycle, with different key dimensions and teams involved.



The Core Functions of AI Risk Management

The steps are performed continuously and iteratively across the AI lifecycle. Govern cuts across all other functions.

Govern

Category	What it covers
1: Policies and practices in place, transparent, and effective.	Legal and regulatory requirements are understood. Trustworthy AI characteristics embedded in org policies. Risk tolerance defined, risk management processes transparent, AI systems inventoried, and decommissioning procedures exist.
2: Accountability structures ensure the right people are empowered and trained.	Roles and responsibilities clearly documented across the org. Personnel receive AI risk training. Executive leadership takes ownership of risk decisions.
3: Diversity, equity, and inclusion prioritized across the AI risk lifecycle.	Risk decisions informed by demographically and disciplinarily diverse teams. Human-AI configurations and oversight responsibilities clearly defined.
4: Culture actively considers and communicates AI risk.	A safety-first mindset embedded in design, development, and deployment. Teams document and communicate risks and impacts. Practices for testing, incident identification, and information sharing in place.
5: Robust engagement with relevant AI actors.	External feedback on individual and societal AI impacts actively collected and integrated. Mechanisms exist to incorporate feedback into system design and implementation.
6: Third-party and supply chain AI risks addressed.	Policies cover IP, rights, and legal risks from third-party entities. Contingency processes exist for failures in high-risk third-party data or AI systems.

Map

Category	What it covers
1: Context is established and understood.	Intended purposes, deployment settings, and context-specific norms documented. A diverse interdisciplinary team engaged. Org mission, business context, risk tolerances, and socio-technical implications all accounted for.
2: The AI system is categorized.	Specific tasks and methods the system supports are defined. Knowledge limits and human oversight mechanisms documented. Scientific integrity and TEVV considerations identified.
3: Capabilities, usage goals, and costs/benefits understood.	Potential benefits and costs, including non-monetary ones, examined. Application scope specified. Processes for operator proficiency and human oversight defined and documented.
4: Risks and benefits mapped for all components including third-party.	AI technology and legal risks across all system components, including third-party software and data, mapped. Internal risk controls identified and documented.
5: Impacts to individuals, groups, and society characterized.	Likelihood and magnitude of both beneficial and harmful impacts documented. Engagement practices for collecting and integrating feedback from affected communities established.

Measure

Category	What it covers
1: Appropriate methods and metrics identified and applied.	Risk metrics selected starting with the most significant risks. Metrics and controls regularly assessed for appropriateness. Independent assessors, not front-line developers, involved in evaluations.
2: AI systems evaluated for trustworthy characteristics.	System tested across all seven trustworthy AI characteristics: validity and reliability, safety, security and resilience, transparency and accountability, explainability, privacy, and fairness. Performance measured against real deployment conditions, with production monitoring in place.
3: Mechanisms for tracking AI risks over time are in place.	Approaches to identify and track existing, unanticipated, and emergent risks established. Risk tracking considered even where formal metrics are unavailable. End user and community feedback and appeal processes integrated into evaluation.
4: Feedback on measurement efficacy gathered and assessed.	Measurement approaches connected to deployment context and informed by domain experts. Results validated by relevant AI actors. Performance improvements and declines identified and documented.

Manage

Category	What it covers
1: AI risks from MAP and MEASURE prioritized, responded to, and managed.	Go/no-go deployment decision made based on whether the system achieves its intended purpose. Risks prioritized by impact and likelihood. Each risk gets a documented response: Mitigate, Transfer, Avoid, or Accept. Residual risks documented.
2: Strategies to maximize benefits and minimize negative impacts planned and documented.	Required resources and non-AI alternatives considered. Mechanisms sustain value of deployed systems. Procedures exist for responding to newly identified risks. Systems that perform inconsistently can be deactivated.
3: Third-party AI risks and benefits managed.	Third-party risks regularly monitored and controls documented. Pre-trained models used in development monitored as part of ongoing system maintenance.
4: Risk treatments, response, recovery, and communication plans documented and monitored.	Post-deployment monitoring plans implemented with mechanisms for user input, incident response, and change management. Continual improvement integrated into system updates. Incidents and errors communicated to all relevant AI actors and recovery processes documented.