2026-04-07

Concept Note:
Artificial Intelligence Risk Management Framework: Trustworthy AI in Critical Infrastructure Profile

Raymond Sheh, Martin Stanley | aiframework@nist.gov
National Institute of Standards and Technology (NIST) Information Technology Laboratory (ITL)

To meet the demand for enhanced safety, security, reliability, capacity, and efficiency, the nation's Critical Infrastructure (CI) will increasingly rely on technological advancements such as Artificial Intelligence (AI) across Information Technology (IT), Operational Technology (OT), and Industrial Control Systems (ICS). Adopting AI in these high-stakes environments relies on AI systems being worthy of trust. The NIST AI Risk Management Framework (AI RMF) was developed to define and promote trustworthiness in AI systems through a repeatable, full lifecycle approach that organizations can use to unlock the benefits of AI while appropriately managing risks.

As part of its Strategy for American Technology Leadership in the 21st Century, the National Institute of Standards and Technology (NIST) Information Technology Laboratory (ITL) is supporting critical infrastructure sectors by launching the development of the AI RMF Trustworthy AI in Critical Infrastructure Profile. This profile will guide CI operators towards specific risk management practices to consider when engaging AI-enabled capabilities. It will then help them to communicate their trustworthiness requirements in an actionable way to teams, developers, and other stakeholders across the AI and CI lifecycles and supply chains.

This profile will address AI trustworthiness characteristics as defined in the NIST AI RMF. Examples of AI systems that may be used in CI, with features that can improve their trustworthiness, include, but are not limited to:

- AI agents for autonomous cybersecurity incident response that include tested, evaluated, validated, and verified guardrails.
- AI-enabled facility and plant monitoring systems that are hardened against adversarial input and monitored for changes in the environment outside verified regions of validity.
- AI-enhanced deterministic diagnostic assistants that utilize AI bills of materials to provide traceable, auditable rationales for recommendations.
- Physics-informed neuro-symbolic AI systems for predicting and maintaining system stability that include verifiable performance guarantees.
- Autonomous robots and vehicles that leverage multimodal sensing, redundant safety systems, and deterministic fail-safe controllers.
- AI-powered digital twins for proactively managing distributed critical data centers to maintain operation during emergencies without overloading fragile utility infrastructure.
- AI optimization systems that degrade gracefully and transparently in response to adverse conditions while alerting human supervisors to take additional measures.
- AI-enabled, transparent, and explainable compliance and risk monitoring systems to improve governance responsiveness while maintaining human-in-the-loop oversight.

This profile aims to align with, contextualize, reference, interpret, adapt, and facilitate the operationalization of existing and upcoming guidance documents at the intersection of AI, IT, OT, ICS, software development, cybersecurity, and critical infrastructure. This profile and associated resources will apply the AI RMF in ways that include, but are not limited to:

- Harmonizing and bridging definitions for key terms and concepts at the intersection of AI, critical infrastructure, and related domains to facilitate efficient, effective cross-sector co-operation and interoperability.
- Guiding requirements analysis to tailor the risk management of AI systems to the performance and reliability expectations and operational realities of critical infrastructure, including legacy systems, physically distributed assets, and resourcing.
- Addressing stringent requirements in the critical infrastructure sector, including the need for deterministic behavior, explainability, graceful degradation, and fail-safe operation.
- Emphasizing the heightened need for adversarial robustness in all lifecycle stages of AI in critical infrastructure.
- Supporting the critical infrastructure needs for rigorous testing, evaluation, validation, and verification (TEVV) of systems, including those that include AI.
- Illuminating critical infrastructure-specific capabilities and trade-offs with competing and complementary AI techniques, technologies, and approaches.
- Promoting visibility and collaboration across the supply chain of AI to address the unique needs, challenges, and risks of AI in critical infrastructure.
- Highlighting practical, actionable, and measurable steps that can be taken by stakeholders at any level of AI expertise and risk management maturity.

To this end, NIST invites all stakeholders to join the NIST community of interest and provide relevant input and feedback via seminars, working sessions, and responses to potential requests for information, position papers, and drafts. Examples of desired information include:

- Current and emerging use cases for AI in critical infrastructure applications.
- Governance challenges unique to the use of AI in critical infrastructure, including those relating to OT, ICS, and cyberphysical systems.
- Existing AI, cybersecurity, and other risk management policies and guidance that need to be reinterpreted to apply meaningfully to the use of AI in critical infrastructure.
- Common questions, points of pain, and sources of confusion, contradiction, and ambiguity relating to the development and adoption of AI in critical infrastructure.
- Other relevant standards, policies, industry conventions, and governance artifacts that the profile should seek to align with.
- Gaps in practical, actionable guidance available to stakeholders from different backgrounds and sectors.

NIST looks forward to engaging with industry, user groups, regulators, policymakers, academia, other stakeholders, and the broader community. Working collaboratively, NIST will develop a profile that provides all critical infrastructure sectors with increased confidence to deploy AI agents and tools as part of their overall strategy, and their developers and vendors with guidance and greater certainty to catalyze and highlight the development of innovative solutions that are based on managing risks and that are worthy of our trust.