MITIGATING ARTIFICIAL INTELLIGENCE (AI) RISK:

Safety and Security Guidelines for Critical Infrastructure Owners and Operators

Publication: April 2024
Department of Homeland Security

Table of Contents

Disclaimer: This document is not meant to bind the public in any way and is only intended to provide clarity to the public. This document is not intended to and does not create any right or benefit, substantive or procedural, enforceable at law or in equity, against the United States, its departments, agencies, or other entities, its officers or employees, or any other person.

EXECUTIVE SUMMARY

The U.S. Department of Homeland Security (DHS) was tasked in *Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*[1] to develop safety and security guidelines for use by critical infrastructure owners and operators. DHS developed these guidelines in coordination with the Department of Commerce, the Sector Risk Management Agencies (SRMAs) for the 16 critical infrastructure sectors, and relevant independent regulatory agencies.

The guidelines begin with insights learned from the Cybersecurity and Infrastructure Security Agency's (CISA) cross-sector analysis of sector-specific AI risk assessments completed by SRMAs and relevant independent regulatory agencies in January 2024.[2] The CISA analysis includes a profile of cross-sector AI use cases and patterns in adoption and establishes a foundational analysis of cross-sector AI risks across three distinct types: 1) Attacks Using AI, 2) Attacks Targeting AI Systems, and 3) Failures in AI Design and Implementation. **DHS drew upon this analysis, as well as analysis from existing U.S. government policy, to develop specific safety and security guidelines to mitigate the identified cross-sector AI risks to critical infrastructure.** The guidelines incorporate the National Institute of Standards and Technology (NIST) AI Risk Management Framework (AI RMF), including its four functions that help organizations address the risks of AI systems: Govern, Map, Measure, and Manage.[3]

While the guidelines in this document are written broadly so they are applicable across critical infrastructure sectors, DHS encourages owners and operators of critical infrastructure to consider sector-specific and context-specific AI risks and mitigations.

[1] Section 4.3(a)(iii) of *Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence* directs DHS as follows: "Within 180 days of the date of this order, the Secretary of Homeland Security, in coordination with the Secretary of Commerce and with SRMAs and other regulators as determined by the Secretary of Homeland Security, shall incorporate as appropriate the AI Risk Management Framework, NIST AI 100-1, as well as other appropriate security guidance, into relevant safety and security guidelines for use by critical infrastructure owners and operators."

[2] The sector-specific risk assessments were developed in response to Section 4.3(a)(i) of Executive Order 14110, which directed SRMAs and independent regulatory agencies, as appropriate, to annually assess the risks related to the use of AI in critical infrastructure.

[3] NIST Artificial Intelligence Risk Management Framework (AI RMF 1.0). See: NIST AI Risk Management Framework.

INTRODUCTION

While artificial intelligence[4] (AI) has the potential to deliver transformative solutions for U.S. critical infrastructure, the introduction of AI systems[5] into critical infrastructure[6] has the potential to make those systems more vulnerable to critical failures, physical attacks, and cyberattacks. At the same time, AI-powered technologies also present new ways for adversaries to expand and enhance attacks on U.S. systems. For owners and operators of critical infrastructure, whose essential services and functions Americans depend on daily, mitigating AI risks is not merely an operational need but a national security and public safety imperative.

In response to *Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence Section 4.3(a)(iii),* the U.S. Department of Homeland Security (DHS) developed these guidelines for use by critical infrastructure owners and operators. **The guidelines address cross-sector AI risks[7] that impact the safety and security of critical infrastructure systems and their functions.** DHS developed these guidelines in coordination with the Department of Commerce, the Sector Risk Management Agencies (SRMAs),[8] and other critical infrastructure sector regulators.

Leveraging the Cybersecurity and Infrastructure Security Agency's (CISA) expertise as National Coordinator for critical infrastructure security and resilience, DHS started with insights learned from the cross-sector analysis of the sector-specific AI risk assessments completed by the SRMAs and relevant independent regulatory agencies in January 2024. DHS also drew insights from other resources including *Guidelines for secure AI system development*,[9] the joint *Cybersecurity Information Sheet Deploying AI Systems Securely*,[10] and the White House Office of Management and Budget (OMB) memorandum on government agency use of AI.[11] DHS then synthesized and framed the subsequent guidelines within the National Institute of Standards and Technology (NIST) AI Risk Management Framework (AI RMF). **DHS selected guidelines that mitigate specific risks identified by the cross-sector analysis and that address AI RMF functions and corresponding subcategories specific to safety and security.**

---

[4] In this document, AI has the meaning set forth in 15 U.S.C. § 9401(3) and in Executive Order 14410 and is defined as "a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. Artificial intelligence systems use machine- and human-based inputs to perceive real and virtual environments; abstract such perceptions into models through analysis in an automated manner; and use model inference to formulate options for information or action."

[5] In this document, "AI system" has the meaning set forth in Executive Order 14410 and is defined as "any data system, software, hardware, application, tool, or utility that operates in whole or in part using AI."

[6] In this document, critical infrastructure has the meaning set forth in 42 U.S.C. . § 5195c(e) and in Presidential Policy Directive 21 (PPD-21) as "systems and assets, whether physical or virtual, so vital to the United States that the incapacity or destruction of such systems and assets would have a debilitating impact on security, national economic security, national public health or safety, or any combination of those matters."

[7] In this document, "risk" has the meaning set forth by NIST as "the composite measure of an event's probability of occurring and magnitude or degree of the consequences of the corresponding event." See: NIST AI Risk Management Framework.

[8] Each critical infrastructure sector has a designated Sector-Specific Agency (SSA) as identified in PPD-21 that coordinates and collaborates with DHS and other relevant Federal departments and agencies to identify vulnerabilities and help mitigate incidents. The 2021 National Defense Authorization Act (NDAA) updated the term "Sector-Specific Agency" to "Sector Risk Management Agency." For more details on how PPD-21 directs SRMAs, please see "Appendix B: Roles, Responsibilities, and Capabilities of Critical Infrastructure Partners and Stakeholders" in the National Infrastructure Protection Plan (NIPP) 2013.

[9] In November 2023, CISA and the United Kingdom's National Cyber Security Centre (NCSC) co-developed *Guidelines for Secure AI System Development*, setting guardrails for developers of AI systems and making security a core requirement of AI system development. This publication is co-sealed by CISA, NCSC and 21 other domestic and international cybersecurity organizations representing 18 countries. To read the full publication, see: Joint Guidelines for Secure AI System Development.

[10] In April 2024, NSA's AI Security Center (AISC) released a joint Cybersecurity Information Sheet outlining how organizations can securely deploy AI systems. This publication is co-sealed by CISA, Federal Bureau of Investigation (FBI), the Australian Signals Directorate's Australian Cyber Security Centre (ASD ACSC), the Canadian Centre for Cyber Security (CCCS), the New Zealand National Cyber Security Centre (NCSC-NZ), and the United Kingdom's National Cyber Security Centre (NCSC-UK). To read the full publication, see: Deploying AI Systems Securely.

[11] In March 2024, OMB issued the first government-wide policy M-24-10 directing agencies to advance AI governance and innovation while managing risks from the use of AI in the Federal Government, particularly those affecting the rights and safety of the public. To read the full memo, see: OMB M-24-10: Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence.

The guidelines **specifically address risks to safety and security, which are uniquely consequential to critical infrastructure.** NIST defines "safety" as a property of a system such that it does not, under defined conditions, lead to a state in which human life, health, property, or the environment is endangered; safety involves reducing both the probability of expected harms and the possibility of unexpected harms. "Security" is defined as resistance to intentional, unauthorized act(s) designed to cause harm or damage to a system.[12]

Because AI risks to critical infrastructure are highly contextual, critical infrastructure owners and operators who use AI-systems should account for their specific circumstances as they use these guidelines. These guidelines do not supersede or replace existing legal requirements for critical infrastructure owners, operators, or regulators.

As AI risks and mitigations evolve and new AI systems and use cases are developed, DHS will continue, as necessary, to update these safety and security guidelines for critical infrastructure owners and operators. DHS will also consider developing additional resources, including playbooks, to assist with the implementation of these guidelines based on evolving technology, updates to the standards landscape (including the NIST AI RMF), input from critical infrastructure risk assessments, input from the AI Safety and Security Board, and other stakeholder feedback.

[12] For definitions of 'safety' and 'security', see NIST's The Language of Trustworthy AI: An In-Depth Glossary of Terms.

AI RISKS TO CRITICAL INFRASTRUCTURE

Any effort to mitigate risk should start with an assessment of those risks. In response to Executive Order 14110 Section 4.3(a)(i), CISA worked with SRMAs and relevant independent regulatory agencies charged with performing annual assessments of AI risks to U.S. critical infrastructure in their respective sectors. The scope of these assessments covered ways in which deploying AI may make critical infrastructure systems more vulnerable to critical failures, physical attacks, and cyberattacks.

This section summarizes the individual sector-specific AI risk assessments to establish a foundational **analysis of cross-sector AI risks**, categorizing risk into three distinct types: 1) Attacks Using AI, 2) Attacks Targeting AI Systems, and 3) Failures in AI Design and Implementation.[13] This section also **profiles cross-sector AI use cases** and **cross-sector patterns in adoption of** AI. Critical infrastructure owners and operators should consider the findings and AI risks detailed in this section when implementing AI safety and security guidelines.

The following **key findings**—drawn from SRMAs' sector submissions—highlight commonalities related to cross-sector AI risks to U.S. critical infrastructure:

- SRMAs consistently highlighted the possibilities of AI as a transformative technology for many critical infrastructure functions; however, they also noted the **tension between the benefits of AI and the risks** introduced by a complex and rapidly evolving technology.
- To date, SRMAs reported their sectors have adopted AI primarily to **support functions that were already partially automated**, and they envision the application of AI to more complex functions as a future advancement.
- SRMAs noted the possibility **that AI could support solutions for many long-standing, persistent challenges**, such as logistics, supply chain management, quality control, physical security, and cyber defense.
- SRMAs consistently viewed AI as a **potential means for adversaries to expand and enhance current cyber tactics, techniques, and procedures.**
- SRMAs identified the following methods to manage and reduce risk to critical infrastructure operations:
    - **Established risk mitigation best practices**, such as information and communications technology (ICT) supply chain risk management, incident response planning, ongoing workforce development, including awareness and training; and
    - **Mitigation strategies more specific to AI,** such as dataset and model validation, human monitoring of automated processes, and AI use policies.

AI Uses and Patterns of Adoption

As part of the sector-specific AI risk assessments, SRMAs identified more than **150 beneficial uses of AI** across their respective sectors. Critical infrastructure owners and operators should use the guidelines in this document to implement AI safely and securely. CISA developed and applied **10 AI use categories** for ease of interpretation and discussion. These AI use categories are likely to evolve in future summaries as more complex applications are introduced to critical infrastructure.

Listed by prevalence as shown in *Figure 1*, these categories include:

[13] As part of the initial 90-day response period defined in Executive Order 14110 and in accordance with CISA's mission to understand, manage, and reduce risk to the nation's cyber and physical infrastructure, CISA identified AI risk categories that align with three of the seven characteristics of AI trustworthiness in the National Institute of Standards and Technology (NIST) AI Risk Management Framework: Safe; Secure & Resilient; and Valid & Reliable. See: NIST AI Risk Management Framework.

- **Operational Awareness**: This involves using AI to gain a clearer understanding of an organization's operations. For instance, AI can be used to monitor network traffic and identify unusual activity, enhancing cybersecurity.
- **Performance Optimization**: This involves using AI to improve the efficiency and effectiveness of processes or systems. For example, AI can be used to optimize supply chain operations, reducing costs, and improving delivery times.
- **Automation of Operations**: This refers to using AI to automate routine tasks and processes in an organization, such as data entry or report generation. For example, AI can be used to automate the process of sorting and analyzing large amounts of data.
- **Event Detection**: This refers to the use of AI to detect specific events or changes in a system or environment. For example, AI can be used in health monitoring systems to detect abnormal heart rates.
- **Forecasting**: This is the use of AI to predict future trends or events based on current and historical data. For instance, AI can be used to forecast sales trends based on past sales data.
- **Research & Development (R&D):** This refers to the use of AI in the development of new products, services, or technologies. For instance, AI can be used in the pharmaceutical industry to expedite the drug discovery process.
- **Systems Planning:** This refers to the use of AI in the planning and design of systems, such as IT infrastructure. For example, AI can be used to predict the performance of a proposed system under various conditions.
- **Customer Service Automation:** This involves using AI to automate aspects of customer service, such as answering frequently asked questions or processing orders. For example, chatbots are a common application of AI in customer service automation.
- **Modeling & Simulation:** This involves using AI to create models and simulations of real-world scenarios. For example, AI can be used to simulate traffic patterns for urban planning purposes.
- **Physical Security:** This refers to the use of AI in maintaining the physical security of a facility or area. For example, AI can be used in surveillance systems to detect intruders or suspicious activity.

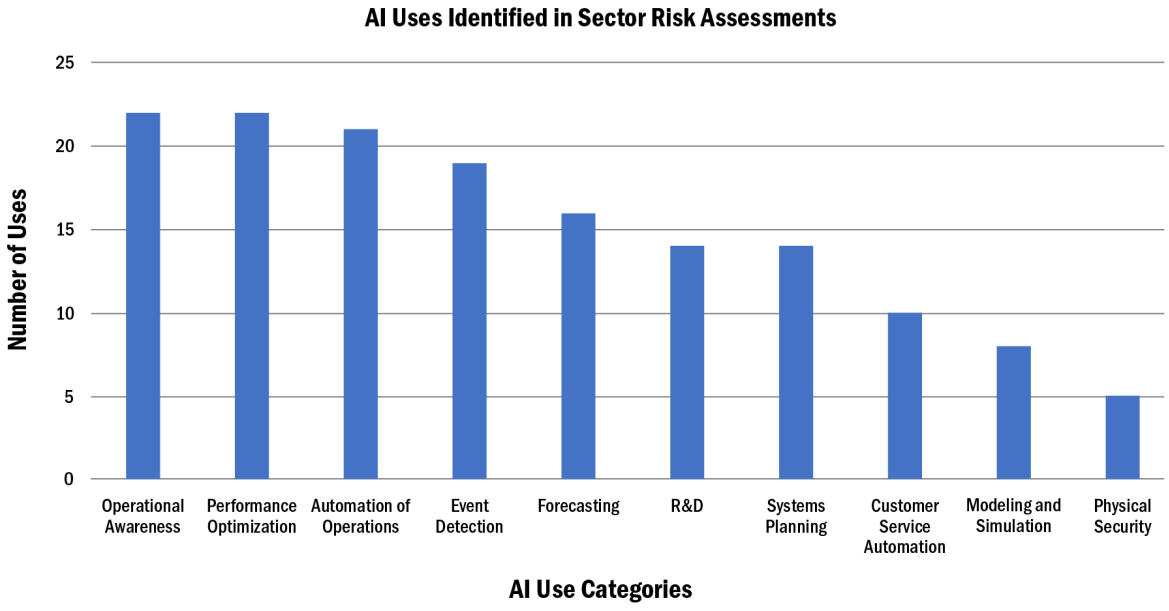

Figure 1: Prevalence of the types of AI uses as reported by the SRMAs in January 2024

SRMAs indicated that the most common critical infrastructure AI use cases involved predictive AI, though recent advances in widely accessible generative AI[14] capabilities may shift that dynamic in future assessments. SRMAs also reported relatively lower levels of current AI adoption for use cases that generate outputs with greater uncertainty or leverage more complex logic, such as forecasting, optimization, modeling, and simulation. This trend is consistent with the overall finding that SRMAs envision the adoption of AI in more complex infrastructure operations as a potential future endeavor for their respective sectors. Finally, most assessments indicated an increasing trend in the degree of AI adoption.

Cross-Sector AI Risk Categories

The guidelines in this document highlight three categories of system-level AI risk, which CISA applied as part of its cross-sector AI risk analysis. In addition to these three categories of AI risk, the cross-sector analysis identified numerous subcategories of AI risk and mitigation strategies. See Appendix A for the full list of risk subcategories and mitigations, as identified by the sectors; all mitigations have also been incorporated in the context of these guidelines. Critical infrastructure owners and operators should consider these **three overarching categories of system-level risk**, as well as sector and context-specific subcategories and mitigations, when implementing the guidelines in this document:

1. **Attacks Using AI**: This risk category refers to the use of AI to automate, enhance, plan, or scale physical attacks on or cyber compromises of critical infrastructure. Common attack vectors include AI-enabled cyber compromises, automated physical attacks, and AI-enabled social engineering.

2. **Attacks Targeting AI Systems**: This risk category largely focuses on targeted attacks on AI systems supporting critical infrastructure. Common attack vectors include adversarial manipulation of AI algorithms, evasion attacks, and interruption of service attacks.

3. **Failures in AI Design and Implementation**: This risk category stems from deficiencies or inadequacies in the planning, structure, implementation, execution, or maintenance of an AI tool or system leading to malfunctions or other unintended consequences that affect critical infrastructure operations. Common methods of design and implementation failure include autonomy, brittleness, and inscrutability.

[14] In this document, generative AI has the meaning set forth in Executive Order 14110 and is defined as: "the class of AI models that emulate the structure and characteristics of input data to generate derived synthetic content. This can include images, videos, audio, text, and other digital content." This is different from AI systems which generate predictions, which NIST defines as "Forecasting quantitative or qualitative outputs through function approximation, applied on input data or measurements." See also: The Language of Trustworthy AI: An In-Depth Glossary of Terms.

GUIDELINES FOR CRITICAL INFRASTRUCTURE OWNERS AND OPERATORS

AI risk management for critical infrastructure is a **continuous process performed throughout the AI lifecycle.**[15] Critical infrastructure owners and operators should consider the three AI risk categories noted above — attacks using AI, attacks on AI, and AI design and implementation failures — when implementing these guidelines.

AI risks are also contextual. Critical infrastructure owners and operators should **account for their own sector-specific and context-specific use of AI** when assessing AI risks and selecting appropriate mitigations. Some mitigations may address multiple risks, while others will be narrowly focused. While the guidelines broadly apply to all sixteen critical infrastructure sectors, specific sectors have already developed[16] and may continue to refine guidelines for managing AI risk tailored to specific settings and contexts and for use as part of annual sector-specific AI risk assessments.

Critical infrastructure owners and operators may focus on different aspects of the AI lifecycle depending on their sector or role. In some cases, critical infrastructure owners and operators will be involved in the design, development, or procurement of AI systems. In other cases, they may not be the original designers or developers of AI systems, but may have a level of responsibility to bear in deploying, operating, managing, maintaining, or retiring these systems.

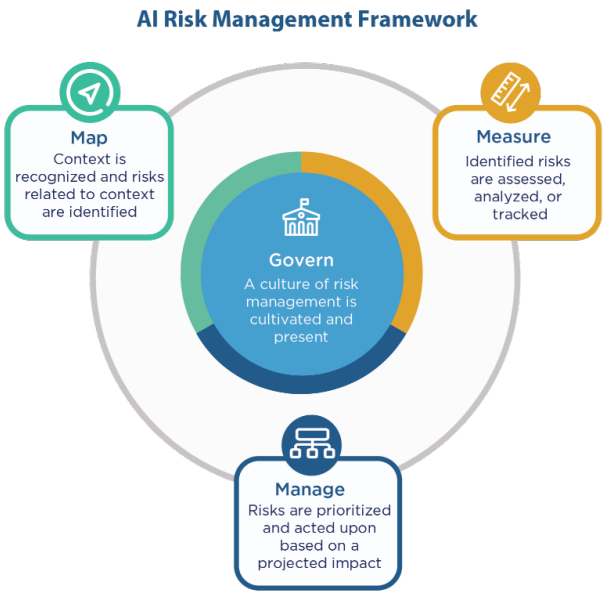

Figure 2: NIST AI Risk Management Framework[1]

In many cases, AI vendors[17] will also play a major role in ensuring the safe and secure use of AI systems for critical infrastructure. Certain guidelines apply both to critical infrastructure owners and operators as well as AI vendors.[18] **Critical infrastructure owners and operators should understand where these dependencies on AI vendors exist and work to share and delineate mitigation responsibilities accordingly.**

These guidelines, which are aligned to the NIST AI RMF, foster efforts to integrate the AI RMF into critical infrastructure enterprise risk management programs. The AI RMF Core, as shown in Figure 2, is comprised of four functions that help organizations address the risks of AI systems: Govern, Map, Measure, and Manage. **Govern** is the cross-cutting AI RMF function that establishes an organizational approach to AI Risk Management as part of the existing Enterprise Risk Management (ERM) function. Recommended actions to be addressed repeatedly, throughout the AI lifecycle, are incorporated in the **Map, Measure,** and **Manage** functions. These guidelines enhance AI safety and security risk management practices contained in the NIST AI RMF. For a recommended mapping of these guidelines to the AI RMF, see Appendix B.

[15] The AI lifecycle referenced in this document is consistent with NIST AI 100-1: AI Risk Management Framework and the OECD Framework for the Classification of AI Systems.
[16] On March 2024, the Department of Treasury published a report identifying the landscape of AI-related cybersecurity and fraud risks in the financial services sector. See: Managing Artificial Intelligence-Specific Cybersecurity Risks in the Financial Services Sector.
[17] In this document, "AI vendor" has the meaning of a commercial supplier of an AI system. NIST defines "vendor" as a commercial supplier of software or hardware, and "AI system" has the meaning set forth in Executive Order 14410 and is defined as "any data system, software, hardware, application, tool, or utility that operates in whole or in part using AI."
[18] For more on the responsibilities of AI vendors, see *Govern D* in the Guidelines for Critical Infrastructure Owners and Operators section below.

Govern: Establish an organizational culture of AI risk management.

The guidelines in this section **support the establishment of policies, processes, and procedures** to anticipate, identify, and manage the benefits and risks of AI at all points in the AI lifecycle. Critical infrastructure owners and operators can foster a culture of risk management by aligning AI safety and security priorities with their own organizational principles and strategic priorities. This organizational approach follows a "secure by design" philosophy where leaders **prioritize and take ownership of safety and security outcomes and build organizational structures that make security a top priority**.[19,20]

a. **Detail plans for cybersecurity risk management**, incident response, security awareness, and safety procedures that include attacks using AI, attacks on AI, and AI design and implementation failures.
b. **Establish secure by design practices** throughout the AI development lifecycle, from design and training to deployment and maintenance, ensuring robustness against attacks and considering AI design and implementation risks.
c. **Establish and track critical AI enterprise data**, including policies and timelines for data retention and disposal.
d. **Establish roles and responsibilities with AI vendors** for the safe and secure operation of AI systems in critical infrastructure contexts. This includes documented plans and regular communication regarding the integration testing, data, input, model and functional validation, and continuous maintenance and monitoring of the AI systems.
e. **Invest in workforce development** and subject matter expertise that establish and sustain a skilled, vetted, and diverse AI workforce.
f. **Assess the advantages and tradeoffs for developing an AI system internally, procuring an AI system, or working with a vendor** to customize an existing AI system.[21]
g. **Establish transparency in AI system use** and accountability mechanisms for AI-driven actions.
h. **Integrate AI threats, AI incidents, and AI system failures into all information-sharing mechanisms** for relevant internal and external stakeholders.
i. **Collaborate with government and industry groups**, such as Sector Coordinating Councils (SCCs), Government Coordinating Councils (GCCs), and Information Sharing and Analysis Centers (ISACs)[22] to inform risk management tools and methodologies.

Map: Understand your individual AI use context and risk profile.

The guidelines in this section establish **the foundational context from which owners and operators of critical infrastructure can evaluate and mitigate AI risks**. Critical infrastructure owners and operators should first understand how, where, and why AI systems will be used, in order to assess context-specific and sector-specific risks and address potential impacts to safety and security.

a. **Inventory all current or proposed AI use cases** in critical infrastructure contexts. Document context-specific and sector-specific AI risks, including risks of: attacks using AI, attacks on AI, and AI design and

[19] For more information see CISA's Principles and Approaches for Secure by Design Software, published jointly with U.S. and international partners.
[20] For more on adopting a holistic process to assessing threats to your AI system, see "Model the threats to your system" guidance in Joint Guidelines for Secure AI System Development (pg. 9)
[21] For more on assessing tradeoffs, see "Design your system for security as well as function and performance" and "Consider security benefits and trade-offs when selecting your AI model" guidance in Joint Guidelines for Secure AI System Development (pg. 9-10).
[22] ISACs help critical infrastructure owners and operators protect their facilities, personnel, and customers from cyber and physical security threats and other hazards. ISACs collect, analyze, and disseminate actionable threat information to their members and provide members with tools to mitigate risks and enhance resiliency. The concept of ISACs was introduced and promulgated pursuant to Presidential Decision Directive-63 (PDD-63), signed May 22, 1998, after which the federal government asked each critical infrastructure sector to establish sector-specific organizations to share information about threats and vulnerabilities. To learn more, see: National Council of ISACs.

implementation failures. Evaluate risk controls or mitigations in place for documented AI risks to AI systems.[23,24]

b. **Document potential context-specific safety and security impacts** associated with AI systems that affect individuals, communities, and society. Include potential risks related to bias, privacy, and misuse of sensitive information.[25]
c. **Require thorough AI impact assessments as part of deployment or acquisition process for AI systems,** documenting the intended purposes, expected benefits, potential safety and security risks, and the quality and appropriateness of the relevant data.
d. **Establish which AI systems in critical infrastructure operations should be subject to human supervision** and human control of decision-making to address malfunctions or unintended consequences that could affect safety or operations.
e. **Review AI vendor supply chains for security and safety risks.** This review should include vendor-provided hardware, software, and infrastructure to develop and host an AI system and, where possible, should incorporate vendor risk assessments and documents, such as software bills of materials (SBOMs), AI system bills of materials (AIBOMs), data cards, and model cards.[26]
f. **Gain and maintain awareness of new failure states** and alternate process redundancy as a result of incorporating AI systems in safety and incident response processes.[27]

Measure: Develop systems to assess, analyze, and track AI risks.

The guidelines in this section identify **repeatable methods and metrics for measuring and monitoring AI risks and impacts** throughout the AI system lifecycle. Critical infrastructure owners and operators can develop their own context-specific testing, evaluation, verification, and validation (TEVV) processes[28] to inform usage and AI risk management decisions.

a. **Define metrics and approaches** for detecting, tracking, and measuring known risks, errors, incidents, or negative impacts.
b. **Continuously test AI systems for errors or vulnerabilities**, including both cybersecurity and compliance vulnerabilities.[29] Use dedicated, segregated networks for testing, research, and development.[30]

---

[23] For more on diagnosing AI-related assets for your organization, see "Identify, track, and protect your assets" guidance in Joint Guidelines for Secure AI System Development (pg. 12).
[24] For more on documentation, see "Document your data, models, and prompts" guidance in Joint Guidelines for Secure AI System Development (pg. 12).
[25] For a definition of "sensitive data'' see: The Language of Trustworthy AI: An In-Depth Glossary of Terms.
[26] For a detailed definition and explanation of SBOMs, see Executive Order 14028 Section 10(j). SBOMs provide an inventory of all open source and proprietary components used to build software. An AIBOM contains similar information for AI systems. SBOMs and AIBOMs provide insight into the supply chain behind software and AI systems allowing for vulnerability analysis to analyze potential risks of a given system. Data and model cards serve a similar purpose, but instead they provide detailed information about the data sets used to train AI systems and the models that power the systems. An adopter of a given AI system can develop a better understanding of what went into the development of a model and its intended use by reviewing the associated data and model cards, if available, and can use this information to assess the safety and security risks of the AI system for their use. The software security and AI communities are working to create more clear guidance around harmonized AIBOMs, potentially integrating data and model cards, that can be generated and consumed by automated tools.
[27] Failure states can be more difficult to identify in AI systems than other types of systems, since AI systems might provide outputs that look and feel accurate but are not actually (e.g., hallucinations).
[28] For more on how testing, evaluation, verification, and validation (TEVV) processes can be applied in an AI context, see NIST's resources for AI Test, Evaluation, Verification, and Validation (TEVV). It is recommended that critical infrastructure owners and operators review NIST AI 200-2 "Guidelines for Evaluating and Red-Teaming Generative AI Models and Systems and Dual Use Foundation Models" once released. This document specifically applies to red-teaming for generative AI and dual-use foundation models.
[29] This might include basic vulnerability scanning processes (see, for example: CISA Vulnerability Scanning) or more targeted penetration testing and red-teaming engagements. AI-powered anomaly detection can also be valuable for identifying unusual patterns in system behavior that might indicate an attack.
[30] For more on protecting your model from indirect or direct access, see "Protect your model continuously" in Joint Guidelines for Secure AI System Development (pg. 14).

c. **Assess performance of risk controls** and mitigations in addressing context-specific AI risks identified during the Map function. Track where mitigations address multiple risk categories or further mitigations are needed to effectively manage a specific risk.
d. **Establish practices for preventing exposure of confidential information** in public AI toolsets, such as private instances for AI models[31] and tools.
e. **Measure AI model performance and outputs** to identify changes in behavior and validate the accuracy of AI systems results post-deployment. Include measures for common AI design and implementation failures, such as brittleness or inscrutability.[32]
f. **Test and evaluate** the context-specific safety and security impacts of AI systems in real-world settings, including through red-teaming exercises.[33]
g. **Evaluate AI vendors and AI vendor systems** from a safety and security standpoint, assessing key areas of concern, such as data drift or model drift, vendor AI expertise, and ongoing operation and maintenance.
h. **Build AI systems with resilience in mind**, enabling quick recovery from disruptions and maintaining functionality under adverse conditions during the deployment phase.[34]
i. **Regularly review risks for which no measure exists or for which measurement is inadequate** to identify gaps and develop newly applicable metrics and measurement approaches.
j. **Establish processes for reporting AI safety and security information** to, and receiving feedback from, potentially impacted communities and stakeholders.[35]

Manage: Prioritize and act upon AI risks to safety and security.

The guidelines in this section **define risk management controls** and best practices for implementing and maintaining them to increase the benefits of AI systems while decreasing the likelihood of harmful safety and security impacts. In order to implement the Manage function well, critical infrastructure owners and operators should regularly allocate resources and apply mitigations, as outlined by governing processes, to mapped and measured AI risks.

a. **Prioritize identified AI safety and security risks** using an evidence-based approach to mitigate potential negative outcomes.
b. **Follow cybersecurity best practices**, including vulnerability management and patching for both software and hardware, using role- and identity-based access controls, and logging and monitoring system access and use.[36]

---

[31] In this document, AI model has the meaning set forth in Executive Order 14110 and is defined as "a component of an information system that implements AI technology and uses computational, statistical, or machine-learning techniques to produce outputs from a given set of inputs."
[32] For more on measuring outputs and performance of your model, see "Monitor your system's behavior" in Joint Guidelines for Secure AI System Development (pg. 16) and "Actively monitor model behavior" in Deploying AI Systems Securely (pg. 7).
[33] In this document, AI red-teaming has the meaning set forth in Executive Order 14110 and is defined as "a structured testing effort to find flaws and vulnerabilities in an AI system, often in a controlled environment and in collaboration with developers of AI. AI red-teaming is most often performed by dedicated 'red teams' that adopt adversarial methods to identify flaws and vulnerabilities, such as harmful or discriminatory outputs from an AI system, unforeseen or undesirable system behaviors, limitations, or potential risks associated with the misuse of the system."; AI red-teaming is still developing as a field of practice and not fully defined, especially due to the rapidly evolving nature of AI technology. As a result, critical infrastructure owners and operators can expect best practices for AI red-teaming to evolve overtime. It is recommended that CI operators and owners review NIST AI 200-2 "Guidelines for Evaluating and Red-Teaming Generative AI Models and Systems and Dual Use Foundation Models" once released. This document specifically applies to red-teaming for generative AI and dual-use foundation models.
[34] For more on deployment governance, see "Manage deployment environment governance" guidance in Deploying AI Systems Securely (pg. 3).
[35] For more on information-sharing, see "Collect and share lessons learned" in Joint Guidelines for Secure AI System Development (pg. 16).
[36] Critical infrastructure owners and operators should also review CISA's Cross-Sector Cybersecurity Performance Goals as a common set of recommended protections for critical infrastructure entities.

c. **Implement new or strengthened mitigation strategies**, where necessary, to address AI risks to safety and security.[37] Mitigation strategies should be risk-informed and context-specific.
d. **Implement tools, such as watermarks,[38] content labels, and authentication techniques**, to assist the public in identifying AI-generated content.[39]
e. **Apply appropriate security controls**, including data integrity checks, encryption, end point protection, data, model and functional validation, data backups, data masking, and defensive AI capabilities to mitigate security risks.[40]
f. **Apply mitigations prior to deployment** of an AI vendor's systems to manage identified safety and security risks and to address existing vulnerabilities, where possible.[41]
g. **Monitor AI systems' inputs and outputs** for unusual or malicious behavior and apply AI behavior analytics for threat detection.[42]
h. **Respond to incidents in accordance with incident management plans** that restore the AI system to safe and secure operation of critical infrastructure systems in the situation where an AI system fails.[43]

[37] An example of a mitigation strategy for AI models is adversarial training. This involves including "adversarial examples" that are added to the AI model training process so that the AI model can recognize and resist certain types of attacks or manipulation. This adversarial training could be informed by learnings from red-teaming and information-sharing.
[38] In this document, watermarking has the meaning set forth in Executive Order 14110 and is defined as: "the act of embedding information, which is typically difficult to remove, into outputs created by AI—including into outputs such as photos, videos, audio clips, or text—for the purposes of verifying the authenticity of the output or the identity or characteristics of its provenance, modifications, or conveyance." The reliability of watermarking schemes is still questionable, with many attacks published on removing and tampering with the embedded watermarks. Red-team assessments could include testing of watermark techniques, to ensure they cannot be spoofed, removed, or bypassed.
[39] In this document, AI-generated content (also termed "synthetic content") has the meaning set forth in Executive Order 14110 and is defined as: "information, such as images, videos, audio clips, and text, that has been significantly modified or generated by algorithms, including by AI."
[40] For more on ensuring AI system's integrity, see "Validate the AI system before and during use" guidance in Deploying AI Systems Securely (pg. 6-7).
[41] For more on applying security best practices prior to deployment, see "Harden deployment environment configurations" guidance in Deploying AI Systems Securely (pg. 4-5).
[42] For more on monitoring inputs, see 'Monitor your system's inputs' guidance in Joint Guidelines for Secure AI System Development (pg. 16).
[43] For more on incident management, see "Develop incident management procedures" guidance in Joint Guidelines for Secure AI System Development (pg. 14).

CONCLUSION

The guidelines in this document address cross-sector AI risks that impact the safety and security of critical infrastructure systems and their functions. Safety and security are uniquely consequential to critical infrastructure and addressing the associated AI risks is not merely an operational need but a national security and public safety imperative.

Although these guidelines are broad enough to apply to all 16 critical infrastructure sectors, AI risks are highly contextual. Therefore, critical infrastructure owners and operators should consider these guidelines within their own specific, real-world circumstances.

As individuals and organizations develop new AI systems and use cases, and as corresponding risks and mitigations evolve, DHS will continue to update these guidelines. DHS will also consider developing additional resources that support critical infrastructure owners and operators in navigating the new opportunities and risks that advances in AI technologies bring in the future.

APPENDIX A: CROSS-SECTOR AI RISKS AND MITIGATION STRATEGIES

The guidelines in this document reflect three categories of system-level AI risk: attacks using AI, attacks on AI, and AI design and implementation failures. This section details CISA's summary of cross-sector AI risk categories, along with risk subcategories and example mitigations.

DHS incorporated each mitigation listed in this appendix in its safety and security guidelines above. For ease of reference, the individual guideline where each risk mitigation is applied is cited in parentheses, e.g., AI RMF category (*Govern, Map, Measure, Manage*) with the unique Guideline ID (*A, B, C,* etc.).

These risk categories and mitigation strategies are based on reported data from individual sectors and **do not contain an exhaustive list of potential or realized sector AI risks**. They are also not listed in priority order (listed out in alphabetical order). These risks and mitigations will likely evolve as new AI systems and use cases are developed.

Risk Category: Attacks Using AI

This risk category covers the use of AI to enhance, plan, or scale physical or cyberattacks on critical infrastructure.

Risk Subcategories – Attacks Using AI:

- **AI-Enabled Cyber Compromises:** Using AI to augment threat actor capabilities and operations (e.g., autonomous malware, reconnaissance, use of deepfakes, automatic parsing of text for vulnerability insights, modeling and model inference and completion, unauthorized data access, cyberattack detection evasion, vulnerability identification and exploitation, machine-speed decision-making, optimization, prompt injections to reveal sensitive information).
- **Automated Physical Attacks:** Attacks on physical infrastructure carried out by autonomous systems (e.g., drone swarms, lethal autonomous weapon systems) either with or without human intervention.
- **Physical Target and Vulnerability Identification:** Using AI systems to collect and analyze data to identify and monitor targets for potential attacks.
- **Social Engineering:** Using AI-enabled psychological manipulation to trick users into revealing sensitive information or performing actions that compromise security controls, including the use of deepfakes or AI-enhanced phishing attempts.
- **Supply Chain Disruptions:** Using AI-enabled attacks, such as cyber compromises, physical attacks, and social engineering to target and disrupt vulnerable logistics supply chains for critical materials.
- **Theft of Intellectual Property and Reverse Engineering:** Using AI systems to collect and interpret public data to reverse engineer intellectual property or other sensitive information.
- **Weapon Development:** Using AI systems to modify or create new weapons or other harmful materials (e.g., improvised explosive devices, chemical or biological weapons) for a physical attack.

Mitigation Strategies – Attacks Using AI:

- **Artificial Intelligence-Generated Content Identification Techniques:** Tools such as watermarking and authentication techniques to assist the public in identifying AI-generated content. *(Manage D)*
- **Defensive Artificial Intelligence Capabilities:** The use of AI to detect, prevent, and respond to physical and cyberattacks against critical infrastructure. *(Manage E)*

- **Encryption:** Protecting data at rest and in transit to maintain its confidentiality through appropriate cryptographic measures, including the implementation of quantum-resistant encryption when appropriate and the development of a cryptographic inventory.[44] *(Manage E)*
- **Host Security:** Detecting, investigating, and responding to threats to protect physical and virtual components of a network from unauthorized access and malicious attacks (e.g., monitoring, patching, sandboxing). *(Govern D; Manage B, G)*
- **Network Security:** The protection of network infrastructure from unauthorized access, misuse, or theft (e.g., access control, network segmentation). *(Govern B, Manage B)*
- **Red-Teaming:** The emulation of a potential adversary's capabilities in a controlled environment in collaboration with AI developers to proactively identify and address vulnerabilities in AI systems, operational processes, and supply chains. *(Measure F)*
- **Secure by Design:** Products that are secure to use out of the box, with little to no configuration changes, are available at no additional cost, and the product's security is a core business requirement and a key consideration during the design phase of a product's development lifecycle (e.g., memory safe programming). *(Govern B, Measure H)*
- **User Security Awareness:** Security practices and foundational training to teach users basic behaviors that help safeguard entities from cyber threats. *(Govern A)*

Risk Category: Attacks on AI

This risk category largely focuses on targeted attacks on AI systems supporting critical infrastructure.[45]

Risk Subcategories – Attacks on AI:

- **Adversarial Manipulation of AI Algorithms or Data:** Intentionally modifying algorithms, data, or sensors to cause AI systems to behave in a way that is harmful to the infrastructure they serve.
- **Evasion Attacks:** The malicious injection of prompts into an AI system to bypass the model, causing system malfunction or the disclosure of sensitive information.
- **Interruption of Service Attacks:** Attacks to render an AI system unavailable to its intended users, either directly (e.g., disrupting the AI system itself) or indirectly (e.g., disrupting the availability of necessary data or computing resources).
- **Loss of Data:** Theft of confidential or sensitive critical infrastructure data from AI systems and other supporting systems.
- **Model Inversion and Extraction:** Malicious attempts to steal the training data or parameters of a model, or reverse engineer the functionality of a model.

Mitigation Strategies – Attacks on AI:

- **Alternate Process Redundancy**: Redundant manual operations with physical device operation, traditional

[44] For more on quantum-resistant encryption, see NIST's Post-Quantum Cryptography project.
[45] Critical infrastructure owners and operators should also review NIST's Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations publication for a more extensive breakdown of AI-specific attacks and mitigations for different classes of AI systems.

computation and analytics, or other manual tasks that normally benefit from high automation. *(Map F)*

- **Data Masking:** Modifying sensitive data in such a way that it is of little to no value to unauthorized intruders while still being usable by software or authorized personnel.[46] *(Manage F)*
- **Dataset Validation:** Efforts to protect the datasets that machine learning[47] and AI algorithms are trained on by filtering poisoned data examples from training, using subject matter expert-annotated datasets, model hardening, two-detector models, and otherwise protecting data from adversary manipulation. *(Govern D, Manage E)*
- **Defensive Artificial Intelligence Capabilities:** The use of AI to prevent, detect, and respond to physical and cyberattacks against critical infrastructure. *(Manage E)*
- **Encryption:** Protecting data at rest and in transit to maintain its confidentiality through appropriate cryptographic measures, including the implementation of quantum-resistant encryption when appropriate and the development of a cryptographic inventory.[48] (*Manage E*)
- **Host Security:** Detecting, investigating, and responding to threats to protect physical and virtual components of a network from unauthorized access and malicious attacks (e.g., monitoring, patching, sandboxing). *(Govern D; Manage B, G)*
- **Human Supervision:** Human oversight in the development, deployment, and operation of AI systems to promote accountability. *(Map D)*
- **ICT Supply Chain Risk Management:** Identifying and managing risks to the information and communications technology (ICT) supply chain, including diversifying ICT supply chain sources across sectors, establishing risk management and contingency plans, and implementing monitoring and response mechanisms to minimize risks to confidentiality, availability, and integrity of sensitive data and critical infrastructure system operations.[49] *(Govern D; Map E; Measure G; Manage F)*
- **Identity and Access Controls:** Limiting access to training data, models, and inputs and controlling system access with strong passwords, multi-factor authentication (MFA), principle of least privilege, and other means to control digital access. *(Manage B)*
- **Input Validation**: Setting strict parameters for system inputs that limit the potential for malicious actors to either disrupt or degrade the AI system. (*Govern D*, *Manage G*)
- **Model Validation**: Efforts to verify machine learning and AI models under development are functioning as intended by assessing the model with a testing data set. *(Govern D; Measure E, G; Manage E)*
- **Network Security:** The protection of network infrastructure from unauthorized access, misuse, or theft (e.g., network segmentation). *(Govern B, Manage B)*
- **Red-Teaming:** The emulation of a potential adversary's capabilities in a controlled environment in collaboration with AI developers to proactively identify and address vulnerabilities in AI systems, operational processes, and supply chains. *(Measure F)*
- **Secure by Design:** Products that are secure to use out of the box with little to no configuration changes, are available at no additional cost, and the product's security is a core business requirement and a key consideration during the design phase of a product's development lifecycle (e.g., memory safe programming).

[46] When it comes to the use of generative AI models, including large language models, a strong security posture includes refraining from use of any sensitive data in prompting models.
[47] In this document, machine learning has the meaning set forth in Executive Order 14110 and is defined as: "a set of techniques that can be used to train AI algorithms to improve performance at a task based on data."
[48] For applications built with generative AI, cryptographic measures would need to be considered across all data layers in the application architecture, including, but not limited to: prompt stores, caches, vector stores, fine-tuning data, and knowledge-bases.
[49] Critical infrastructure owners and operators should also review CISA's ICT Supply Chain Resource Library.

(Govern B, Measure H)

- **Software Bill of Materials:** A formal inventory of software components and dependencies, information about those components, and their hierarchical relationships accompanied by vulnerability exchange files to address software-specific vulnerabilities. This inventory should apply to AI and non-AI systems. *(Map E)*
- **User Security Awareness:** Security practices and foundational trainings to teach users basic behaviors that can help safeguard entities from cyber threats. *(Govern A)*

Risk Category: AI Design and Implementation Failures

This risk category stems from deficiencies or inadequacies in the planning, structure, implementation, or execution of an AI tool or system leading to malfunctions or other unintended consequences that affect critical infrastructure operations.

Risk Subcategories – AI Design and Implementation Failures:

- **Autonomy:** Malfunction or unexpected behavior facilitated by excessive permissions or poorly defined operational parameters for AI systems.
- **Brittleness:** Unintended failure or unexpected behavior of AI systems when confronted with circumstances outside of its original problem context, leading to a lack of robustness.
- **Inscrutability:** Limited interpretability, lack of transparency, or documentation in AI system development or deployment, or inherent uncertainties in AI systems that make diagnosing and correcting AI system anomalies difficult.[50]
- **Inadvertent Systemic and Design Flaws**: Unintentional defects in the design or development of AI systems or models that can lead to unexpected or harmful behavior that disrupts critical infrastructure operations, supply chains, or creates vulnerabilities that adversaries could exploit to do the same.
- **Inconsistent System Maintenance**: Failure to regularly update and maintain AI models and supporting systems, potentially leading to malfunction of service disruptions.
- **Interoperability and Configuration Between AI Systems and Non-AI Systems:** The integration of AI systems into broader IT networks that include non-AI functions could lead to interoperability and compatibility challenges. Additionally, integrating multiple, separate AI systems into a larger network could lead to compatibility challenges between multiple AI models.
- **Overreliance on Artificial Intelligence**: Human operators' excessive reliance on an AI system's ability to make decisions or perform tasks, potentially resulting in operational disruptions if the AI system fails.
- **Statistical Bias**: The reproduction or amplification of computational errors or distortions due to data integrity failures or other design defects. This could result in biased outputs and erroneous decision-making.
- **Subject Matter Expert Shortages:** A shortage of personnel trained in the design, integration, training, management, and interpretation of AI systems could result in inappropriate selection, installation, and use of AI systems.
- **Supply Chain Vulnerabilities:** Third party vendors could use AI products that depend on data sets that have not been validated or other external factors that could lead to malfunction or operational disruption.

[50] One method of addressing concerns around inscrutability is through implementing "explainable AI" systems which are transparent and understandable in their decision-making processes.

- **Under Reliance on Artificial Intelligence**: The insufficient incorporation or use of AI in a system or process that leads to risks or vulnerabilities that the use of AI could prevent or mitigate.

Mitigation Strategies – AI Design and Implementation Failures:

- **Data, Model, and Functional Validation:** Efforts to verify machine learning and AI models under development are functioning as intended by assessing the model with a validated testing data set, including by third-party auditors or other external certification entities. *(Govern D; Measure A, E, G; Manage E)*
- **Human Supervision:** Human oversight in the development, deployment, and operation of AI systems to promote accountability. *(Map D)*
- **ICT Supply Chain Risk Management:** Identifying and managing risks to the ICT supply chain, including diversifying ICT supply chain sources across sectors, establishing risk management and contingency plans, and implementing monitoring and response mechanisms to minimize risks to confidentiality, availability, and integrity of sensitive data and critical infrastructure system operations. *(Govern D; Map E; Measure G; Manage F)*
- **Identity and Access Controls:** Limiting access to training data, models, and inputs and controlling system access with strong passwords, MFA, principle of least privilege, and other means to control digital access. *(Manage B)*
- **Non-public Instances of AI Tools and Models**: Implementation of private instances for AI models and tools to limit and prevent exposure of confidential information in public AI toolsets. *(Measure D)*
- **Software Bill of Materials**: A formal inventory of software components and dependencies, information about those components, and their hierarchical relationships accompanied by vulnerability exchange files to address software-specific vulnerabilities. This inventory should apply to AI and non-AI systems. *(Map E)*

General Mitigations for AI Risks

The following mitigation strategies are broadly applicable to risks from all three risk categories: attacks using AI, attacks on AI, and AI design and implementation failures.

General Mitigation Strategies:

- **Artificial Intelligence Principles of Use:** Federal efforts and industry best practices for AI system development life cycle management that promotes accountability, transparency, reliability, traceability, and governability to support the resilience, robustness, and trustworthiness of AI systems. *(Govern A, B, I; Measure A, G, J; Manage B, D, H)*
- **Building Operational Resiliency:** The creation and implementation of backup systems (including manual and non-AI systems), continuity of operations plans, crisis exercises, and adequate liquidity to maintain operational resiliency and continuity during an event. *(Map F; Measure H; Manage E, H)*
- **Data Inventory:** A detailed inventory of an enterprise's sensitive data, including the timeline over which information should be protected. *(Govern C)*
- **Data Backups:** Maintenance of offline data backups for all AI systems to help ensure continuity of operations during system outages or malfunction. *(Manage E)*
- **End Point Security:** Protective measures that can detect and stop intrusions at vulnerable endpoints from external threats such as malware. *(Manage E)*

- **Cyber Incident Response:** The process through which an organization responds to cyber threats, including preparation, detection and analysis, containment, eradication and recovery, and post-incident activity. Include a response plan specific to AI failure, including how to identify and report problems with an AI tool. *(Govern A, H; Manage H)*
- **Employee Vetting:** Screening people who work in sensitive facilities or with sensitive information to identify those with ties to adversarial groups or other indicators suggesting harmful intent. *(Govern A, E)*
- **Guiding Development of Artificial Intelligence:** Financial investment in AI safety, standards development, and testing by the federal government and AI system owners and operators to strengthen and further the development of AI. *(Govern I)*
- **Information Gathering and Analysis:** Public and private sector efforts to collect and report information on physical and cyber threats for analysis and threat intelligence production. *(Govern H, I; Measure J)*
- **Information Sharing Between Government Agencies:** Communication between all levels of government to enhance threat detection, prevention, and response to physical and cyber threats. *(Measure J)*
- **Information Sharing Between Public and Private Sector:** Communication of physical and cyber threat intelligence between government agencies and private sector critical infrastructure owners and operators. *(Measure J)*
- **Internal Reviews:** Continual evaluation and assessment by organizations to identify vulnerabilities in the software they use. *(Manage B)*
- **Maintaining Public Confidence in AI:** Maintaining confidence in critical infrastructure owners and operators' ability to manage AI systems responsibly and respond to threats and incidents quickly. *(Govern G, H; Map B, D; Measure J)*
- **Model Risk Management:** The identification, assessment, monitoring, and mitigation of risks to AI models that could arise due to issues with data quality, model design, deployment, operation, and decommissioning. *(Govern D; Measure E, G; Manage E)*
- **Non-Public Instances of AI Tools and Models**: Implementation of private instances for AI models and tools to limit and prevent exposure of confidential information in public AI toolsets. *(Measure D)*
- **Preserving Institutional Knowledge**: The policies and capabilities organizations develop to capture and maintain collective expertise, skills, and experience amassed over time including the ability to operate critical systems without AI software. *(Govern E)*
- **Public Sector Incident Response Plans and Implementation:** The creation of public sector incident response plans that include designated roles and responsibilities for emergency responders and law enforcement personnel in the event of a physical or cyber emergency or threat that could affect national security, the economy, or public health and safety. *(Govern A, H)*
- **Research, Development, Monitoring, and Testing Environments:** Dedicated, segregated networks in which critical infrastructure organizations can conduct research and development, and test new AI solutions and monitor implementations for errors or vulnerabilities. (*Measure B)*
- **Vulnerability Management:** Proactive identification, disclosure, and remediation of vulnerabilities in AI models and related software. *(Manage B)*
- **Workforce Development:** Providing AI training to employees to be able to identify any issues in their organizations' workflows. *(Govern E)*

APPENDIX B: GUIDELINES MAPPED TO NIST AI RMF

The guidelines in this document align to NIST AI Risk Management Framework (RMF); critical infrastructure owners and operators can use these guidelines to incorporate RMF elements into their enterprise risk management programs. The AI RMF Core is comprised of four functions that help organizations address the risks of AI systems: Govern, Map, Measure, and Manage. The AI RMF also details subcategories – suggested actions for managing risks - for each these four functions.

This section overlays the guidelines in this document with specific subcategories of the AI RMF (e.g., Govern 1.1, Map 2.3) for which they correspond most closely. This is intended to provide critical infrastructure owners and operators a better **understanding of how specific guidelines may apply when using the AI RMF**.

The NIST AI RMF covers a wide range of topics and risks. In scoping the guidelines for this document, DHS incorporated all AI RMF subcategories related to safety, security, and resilience, as well as other relevant sections.

Govern

DHS Guideline	Corresponding AI RMF Subcategory
A. Detail plans for cybersecurity risk management, incident response, security awareness, and safety procedures that include attacks using AI, attacks on AI, and AI design and implementation failures.	*Govern 1.2* The characteristics of trustworthy AI are integrated into organizational policies, processes, and procedures.
B. Establish secure by design practices throughout the AI development lifecycle, from design and training to deployment and maintenance, ensuring robustness against attacks and considering AI design and implementation risks.	*Govern 1.5* The characteristics of trustworthy AI are integrated into organizational policies, processes, and procedures.
C. Establish and track critical AI enterprise data, including policies and timelines for data retention and disposal.	*Govern 1.6* Mechanisms are in place to inventory AI systems and are resourced according to organizational risk priorities.
D. Establish roles and responsibilities with AI vendors for the safe and secure operation of AI systems in critical infrastructure contexts. This includes documented plans and regular communication regarding the integration testing, data, input, model and functional validation, and continuous maintenance and monitoring of the AI systems.	*Govern 2.1* Roles and responsibilities and lines of communication related to mapping, measuring, and managing AI risks are documented and are clear to individuals and teams throughout the organization.
E. Invest in workforce development that establish and sustain skilled, vetted, and diverse AI workforce.	*Govern 3.1* Decision-makings related to mapping, measuring, and managing AI risks throughout the lifecycle is

	informed by a diverse team (e.g., diversity of demographics, disciplines, experience, expertise, and backgrounds).
F. Assess the advantages and tradeoffs for developing an AI system internally, procuring an AI system, or working with a vendor to customize an existing AI system	*Govern 4.1* Organizational policies, and practices are in place to foster a critical thinking and safety-first mindset in the design, development, deployment, and uses of AI systems to minimize negative impacts.
G. Establish transparency in AI system use and accountability mechanisms for AI-driven actions.	*Govern 4.1* Organizational policies, and practices are in place to foster a critical thinking and safety-first mindset in the design, development, deployment, and uses of AI systems to minimize negative impacts.
H. Integrate AI threats, AI incidents, and AI system failures into all information-sharing mechanisms for relevant internal and external stakeholders.	*Govern 4.3* Organizational practices are in place to enable AI testing, identification of incidents, and information sharing.
I. Collaborate with government and industry groups such as Sector Coordinating Councils (SCCs), Government Coordinating Councils (GCCs), and Information Sharing and Analysis Centers (ISACs) to inform risk management tools and methodologies.	*Govern 5.1* Organizational policies and practices are in place to collect, consider, prioritize, and integrate feedback from those external to the team that developed or deployed the AI system regarding the potential individual and societal impacts related to AI risks.

Map

DHS Guideline	Corresponding AI RMF Subcategory
A. Inventory all current or proposed AI use cases in critical infrastructure contexts. Document context-specific and sector-specific AI risks, including risks of: attacks using AI, attacks on AI, and AI design and implementation failures. Evaluate risk controls or mitigations in place for documented AI risks to AI systems	*Map 1.1* Intended purpose, potentially beneficial uses, context-specific laws, norms and expectations, and prospective settings in which the AI system will be deployed are understood and documented.

B. Document potential context-specific safety and security impacts associated with AI systems that affect individuals, communities, and society. Include potential risks related to bias, privacy, and misuse of sensitive information.	*Map 2.1* The specific task, and methods used to implement the task, that the AI system will support is defined.
C. Require thorough AI impact assessments as part of deployment or acquisition process for AI systems, documenting the intended purposes, expected benefits, potential safety and security risks, and the quality and appropriateness of the relevant data.	*Map 2.1* The specific task, and methods used to implement the task, that the AI system will support is defined.
D. Establish which AI systems in critical infrastructure operations should be subject to human supervision and human control of decision-making to address malfunctions or unintended consequences that could affect safety or operations.	*Map 3.5* Processes for human oversight are defined, assessed, and documented in accordance with organizational policies from the Govern function.
E. Review AI vendor supply chains for security and safety risks. This review should include vendor-provided hardware, software, and infrastructure to develop and host an AI system and, where possible, should incorporate vendor risk assessments and documents, such as software bills of materials (SBOMs), AI system bills of materials (AIBOMs), data cards, and model cards.	*Map 4.1* Approaches for mapping AI technology and legal risks of its components – including the use of third-party data or software – are in place, followed, and documented, as are risks of infringement of a third-party's intellectual property or other rights. *Map 4.2* Internal risk controls for components of the AI system including third-party AI technologies are identified and documented.
F. Gain and maintain awareness of new failure states and alternate process redundancy as a result of incorporating AI systems in safety and incident response processes.	*Map 5.2* Practices and personnel for supporting regular engagement with relevant AI actors and integrating feedback about positive, negative, and unanticipated impacts are in place and documented.

Measure

DHS Guideline	Corresponding AI RMF Subcategory
A. Define metrics and approaches for detecting, tracking, and measuring known risks, errors, incidents, or negative impacts.	*Measure 1.1* Approaches and metrics for measurement of AI risks enumerated during the Map function are selected for implementation starting with the most significant AI risks. The risks or trustworthiness characteristics that will not – or cannot – be measured are properly documented.
B. Continuously test AI systems for errors or vulnerabilities, including both cybersecurity and compliance vulnerabilities. Use dedicated, segregated networks for testing, research, and development.	*Measure 1.1* Approaches and metrics for measurement of AI risks enumerated during the Map function are selected for implementation starting with the most significant AI risks. The risks or trustworthiness characteristics that will not – or cannot – be measured are properly documented. *Measure 2.1* Test sets, metrics, and details about the tools used during test, evaluation, validation, and verification (TEVV) are documented.
C. Assess performance of risk controls and mitigations in addressing context-specific AI risks identified during the Map function. Track where mitigations address multiple risk categories or further mitigations are needed to effectively manage a specific risk.	*Measure 1.2* Appropriateness of AI metrics and effectiveness of existing controls is regularly assessed and updated including reports of errors and impacts on affected communities.
D. Establish practices for preventing exposure of confidential information in public AI toolsets, such as private instances for AI models and tools.	*Measure 2.1* Test sets, metrics, and details about the tools used during test, evaluation, validation, and verification (TEVV) are documented.
E. Measure AI model performance and outputs to identify changes in behavior and validate the accuracy of AI systems results post-deployment. Include measures for common AI design and implementation failures, such as brittleness or inscrutability.	*Measure 2.3* AI system performance or assurance criteria are measured qualitatively or quantitatively and demonstrated for conditions similar to deployment setting(s). Measures are documented.

	Measure 4.3 Measurable performance improvements or declines based on consultations with relevant AI actors including affected communities, and field data about context-relevant risks and trustworthiness characteristics, are identified and documented.
F. Test and evaluate the context-specific safety and security impacts of AI systems in real-world settings, including through red-teaming exercises.	*Measure 2.6* AI system is evaluated regularly for safety risks – as identified in the Map function. The AI system to be deployed is demonstrated to be safe, its residual negative risk does not exceed the risk tolerance, and can fail safely, particularly if made to operate beyond its knowledge limits. Safety metrics implicate system reliability and robustness, real-time monitoring, and response times for AI system failures. *Measure 2.7* AI system security and resilience – as identified in the Map function – are evaluated and documented.
G. Evaluate AI vendors and AI vendor systems from a safety and security standpoint, assessing key areas of concern such as data drift or model drift, vendor AI expertise, and ongoing operation and maintenance.	*Measure 2.6* AI system is evaluated regularly for safety risks – as identified in the Map function. The AI system to be deployed is demonstrated to be safe, its residual negative risk does not exceed the risk tolerance, and can fail safely, particularly if made to operate beyond its knowledge limits. Safety metrics implicate system reliability and robustness, real-time monitoring, and response times for AI system failures. *Measure 2.7* AI system security and resilience – as identified in the Map function – are evaluated and documented.
H. Build AI systems with resilience in mind, enabling quick recovery from disruptions and maintaining functionality under adverse conditions during the deployment phase.	*Measure 2.7* AI system security and resilience – as identified in the Map function – are evaluated and documented.

I. Regularly review risks for which no measure exists or for which measurement is inadequate to identify gaps and develop newly applicable metrics and measurement approaches.	*Measure 3.2* Risk tracking approaches are considered for settings where AI risks are difficult to assess using currently available measurement techniques or where metrics are not yet available.
J. Establish processes for reporting AI safety and security information to, and receiving feedback from, potentially impacted communities and stakeholders	*Measure 4.2* Measurement results regarding AI system trustworthiness in deployment context(s) and across AI lifecycle are informed by input from domain experts and other relevant AI actors to validate whether the system is performing consistently as intended. Results are documented.

Manage

DHS Guideline	Corresponding AI RMF Subcategory
A. Prioritize identified AI safety and security risks using an evidence-based approach to mitigate potential negative outcomes.	*Manage 1.2* Treatment of documented AI risks is prioritized based on impact, likelihood, or available resources or methods.
B. Follow cybersecurity best practices, including vulnerability management and patching for both software and hardware, using role- and identity-based access controls, and logging and monitoring system access and use.	*Manage 1.2* Treatment of documented AI risks is prioritized based on impact, likelihood, or available resources or methods.
C. Implement new or strengthened mitigation strategies, where necessary, to address AI risks to safety and security. Mitigation strategies should be risk-informed and context-specific.	*Manage 1.2* Treatment of documented AI risks is prioritized based on impact, likelihood, or available resources or methods.
D. Implement tools, such as watermarks, content labels, and authentication techniques, to assist the public in identifying AI-generated content.	*Manage 1.3* Responses to the AI risks deemed high priority as identified by the Map function, are developed, planned, and documented. Risk response options can include mitigating, transferring, avoiding, or accepting.
E. Apply appropriate security controls, including data integrity checks, encryption, end point protection, data, model and functional validation, data backups, data masking, and defensive AI capabilities to mitigate security risks.	*Manage 2.2* Mechanisms are in place and applied to sustain the value of deployed AI systems.

F. Apply mitigations prior to deployment of an AI vendor's systems to manage identified safety and security risks and to address existing vulnerabilities, where possible.	*Manage 3.1* AI risks and benefits from third-party resources are regularly monitored, and risk controls are applied and documented.
G. Monitor AI systems' inputs and outputs for unusual or malicious behavior and apply AI behavior analytics for threat detection.	*Manage 4.1* Post-deployment AI system monitoring plans are implemented, including mechanisms for capturing and evaluating input from users and other relevant AI actors, appeal and override, decommissioning, incident response, recovery, and change management.
H. Respond to incidents in accordance with incident management plans that restore the AI system to safe and secure operation of critical infrastructure systems in the situation where an AI system fails.	*Manage 4.1* Post-deployment AI system monitoring plans are implemented, including mechanisms for capturing and evaluating input from users and other relevant AI actors, appeal and override, decommissioning, incident response, recovery, and change management.