U.S. Department of Homeland Security

# Roles and Responsibilities Framework for Artificial Intelligence in Critical Infrastructure

*In Consultation with*

The Artificial Intelligence Safety and Security Board

November 14, 2024

Homeland Security

AI SAFETY & SECURITY BOARD

# Contents

# Letter from the Secretary

America's critical infrastructure systems – systems that power our homes and businesses, deliver clean water, facilitate the digital networks that connect us, and much more – are vital to domestic and global safety and stability. Disruptions to the smooth operation of our nation's hospitals and emergency services, electricity substations, pipelines, water treatment facilities, and other critical systems can have devastating consequences for our security.

Artificial intelligence (AI) is already altering the way Americans interface with critical infrastructure. New technology, for example, is helping to sort and distribute mail to American households, quickly detect earthquakes and predict aftershocks, and prevent blackouts and other electric-service interruptions. These uses do not come without risk, though: a false alert of an earthquake can create panic, and a vulnerability introduced by a new technology may risk exposing critical systems to nefarious actors.

Meanwhile, our adversaries are using and will continue to use AI to launch complex, sophisticated, and frequent cyberattacks on U.S. critical infrastructure. To counter these threats, critical infrastructure operators can use AI to both more effectively defend against malicious attacks and improve the overall resilience of critical infrastructure services.

America's continued security and prosperity will depend on how critical infrastructure stakeholders develop and deploy AI.

The Department of Homeland Security (DHS) is charged with safeguarding America's critical infrastructure in our evolving and expanding threat environment. To inform our approach to this vital task, earlier this year I convened some of the nation's foremost leaders in technology, industry, civil rights, academia, and government to form a first-of-its-kind AI Safety and Security Board. In close consultation with the Board, DHS developed the Roles and Responsibilities Framework for Artificial Intelligence in Critical Infrastructure.

This Framework proposes specific actions that entities across the AI ecosystem can consider, adapt, and implement based on their role and relationship to the critical services that ultimately serve the American people. It is designed to be incorporated into new and existing processes that drive product development, procurement decisions, and information exchanges. It also captures responsibilities that will never be fully complete – such as supporting foundational research in AI safety and security and exploring AI use cases that benefit the American public – but nevertheless require a concerted, national effort to ensure continued progress.

Addressing the extraordinary scale and impact of AI in U.S. critical infrastructure calls for a whole-of-nation approach. This Department has a critical role to play in socializing and harmonizing core elements of this Framework, which draws from the important work of the White House, the AI Safety Institute, and our Cybersecurity and Infrastructure Security Agency. Going forward, we will engage with our critical infrastructure partners to understand how this Framework can be adapted for sector-specific needs. We will also convene dialogues with international partners on how we

can harmonize our approach to AI safety and security across critical infrastructure globally.

As the Secretary of Homeland Security and Chair of the AI Safety and Security Board, I am proud of what we have achieved together. We welcome continued dialogue on these important issues, and we look forward to updating this Framework as this extraordinary technology continues to grow and evolve.

Alejandro N. Mayorkas
Secretary, U.S. Department of Homeland Security
Chair, Artificial Intelligence Safety and Security Board

# Artificial Intelligence Safety and Security Board Membership

The Artificial Intelligence Safety and Security Board (the "Board") advises the Secretary, the critical infrastructure community, other private sector stakeholders, and the broader public on the safe, secure, and responsible development and deployment of AI technology in our nation's critical infrastructure. The duties of the Board are solely advisory in nature. Board Members serve as uncompensated representatives of their sectors. The Board provided advice, information, and recommendations in the development of the Department of Homeland Security's Roles and Responsibilities Framework for AI in Critical Infrastructure (the "Artificial Intelligence Roles and Responsibilities Framework").

**Alejandro N. Mayorkas,** Secretary, U.S. Department of Homeland Security; Chair, Artificial Intelligence Safety and Security Board
**Sam Altman,** CEO, OpenAI
**Dario Amodei,** CEO and Co-Founder, Anthropic
**Ed Bastian,** CEO, Delta Air Lines
**Marc Benioff,** Chair and CEO, Salesforce
**Rumman Chowdhury, Ph.D.,** CEO, Humane Intelligence
**Matt Garman,** CEO, Amazon Web Services
**Alexandra Reeve Givens,** President and CEO, Center for Democracy and Technology
**Bruce Harrell,** Mayor of Seattle, Washington; Chair, Technology and Innovation Committee, United States Conference of Mayors
**Damon T. Hewitt,** President and Executive Director, Lawyers' Committee for Civil Rights Under Law
**Vicki Hollub,** President and CEO, Occidental Petroleum
**Jensen Huang,** President and CEO, NVIDIA
**Arvind Krishna,** Chairman and CEO, IBM
**Fei-Fei Li, Ph.D.,** Co-Director, Stanford Human-centered Artificial Intelligence Institute
**Wes Moore,** Governor of Maryland
**Satya Nadella,** Chairman and CEO, Microsoft
**Shantanu Narayen,** Chair and CEO, Adobe
**Sundar Pichai,** CEO, Alphabet
**Arati Prabhakar,** Ph.D., Assistant to the President for Science and Technology; Director, the White House Office of Science and Technology Policy
**Chuck Robbins,** Chair and CEO, Cisco; Chair, Business Roundtable
**Lisa Su,** Chair and CEO, Advanced Micro Devices (AMD)
**Nicol Turner Lee,** Ph.D., Senior Fellow and Director of the Center for Technology Innovation, Brookings Institution
**Kathy Warden,** Chair, CEO, and President, Northrop Grumman
**Maya Wiley,** President and CEO, The Leadership Conference on Civil and Human Rights

AI SAFETY & SECURITY BOARD

# Executive Summary

Artificial Intelligence (AI) systems are transforming U.S. critical infrastructure, unlocking new opportunities, and presenting new risks to vital systems and services. The choices that organizations and individuals make regarding how AI systems are developed, how they can be accessed, and how they function within larger systems will determine the impact that AI will have when deployed to broad segments of U.S. critical infrastructure. To inform those choices, the Department of Homeland Security (DHS) developed the Roles and Responsibilities Framework for AI in Critical Infrastructure (the "Framework") in close consultation with the Artificial Intelligence Safety and Security Board (the "Board"). The Framework recommends key roles and responsibilities for the safe and secure development and deployment of AI in U.S. critical infrastructure.

This Framework seeks to complement and advance the AI safety and security best practices established by the White House Voluntary Commitments, the Blueprint for an AI Bill of Rights, Executive Order 14110 on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, the OMB M-24-10 Memorandum on Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence, the Memorandum on Advancing the United States' Leadership in Artificial Intelligence, the work of the AI Safety Institute, the DHS Safety and Security Guidelines for Critical Infrastructure Owners and Operators, and others.

The roles, responsibilities, use cases, and risks associated with AI in critical infrastructure are complex, interdependent, and will change over time with the evolution of the technology. Given these considerations, this Framework:

- Proposes a set of voluntary responsibilities for the safe and secure use of AI in U.S. critical infrastructure, divided among five key roles: cloud and compute infrastructure providers, AI developers, critical infrastructure owners and operators, civil society, and the public sector.
- Evaluates these roles across five responsibility areas: securing environments, driving responsible model and system design, implementing data governance, ensuring safe and secure deployment, and monitoring performance and impact for critical infrastructure.
- Provides technical and process recommendations to enhance the safety, security, and trustworthiness of AI systems deployed across the nation's sixteen critical infrastructure sectors.

If adopted and implemented throughout the AI ecosystem, this Framework intends to further AI safety and security in critical infrastructure, including the harmonization of safety and security practices, improve the delivery of critical services, enhance trust and transparency among entities, protect civil rights and civil liberties, and advance AI safety and security research that will further enable critical infrastructure to responsibly operationalize and deploy AI.

This Framework builds upon existing risk frameworks that enable entities to evaluate whether the use of AI for certain systems or applications could result in harms to critical infrastructure assets, sectors, nationally significant systems, or individuals served by such systems. The responsibilities in this Framework have been tailored to address these potential harms through the implementation of technical risk mitigations, accountability mechanisms, routine testing practices, and incident response planning. Importantly, the Framework also prioritizes the role of transparency, communication, and information sharing as key elements of AI safety and security. This Framework does not represent a complete list of the relevant entities' responsibilities for AI safety and security, but instead focuses on activities to advance the safe and secure use

of AI that are particularly relevant to critical infrastructure. All relevant entities, from cloud and compute infrastructure providers to AI developers and critical infrastructure owners and operators, should consider sector-specific and context-specific AI risk mitigations in addition to the foundational responsibilities presented by this Framework.

AI remains an emerging technology, and AI safety and security practices continue to develop in tandem. DHS and the AI Safety and Security Board will lead by example to support, strengthen, and advance the core concepts in this Framework as a shared commitment to protect and shape the future of American innovation in our nation's critical infrastructure.

# I. Introduction

American critical infrastructure encompasses the sixteen sectors of American society whose systems are so vital that their incapacitation would have a debilitating impact on the life of the nation.[1] It includes the U.S. systems of defense, energy, transportation, information technology, financial services, food and agriculture, communications, and others.

With its partners, the U.S. Department of Homeland Security (DHS) helps to protect the methods by which Americans power their homes and businesses, make financial transactions, share information, access and deliver healthcare, and put food on the table – among many other daily activities. AI can be a powerful force to improve the services that critical infrastructure provides, build resilience, detect threats, and support disaster recovery. As the entities that own and operate these critical infrastructure systems increasingly adopt AI, it is the Department's duty to understand, anticipate, and address risks that could negatively affect these systems and the consumers they serve. This duty includes ensuring that critical infrastructure systems function for and serve all people in the United States. AI systems in critical infrastructure are no exception – they must be intentionally designed, developed, and deployed in a manner that is effective and respects and preserves privacy, civil rights, and civil liberties.[2] The Department will do its part to protect these rights as a necessary condition of driving safe and secure outcomes in critical infrastructure.

## The 16 U.S. Critical Infrastructure Sectors

| | | | |
|---|---|---|---|
| Chemical | Commercial Facilities | Communications | Critical Manufacturing |
| Dams | Defense Industrial Base | Emergency Services | Energy |
| Financial Services | Food and Agriculture | Government Facilities | Healthcare and Public Health |
| Information Technology | Nuclear Reactors, Materials, and Waste | Transportation Systems | Water and Wastewater |

Recognizing the critical nature of these systems and their role in using AI to serve the American people, the President's October 2023 Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (Executive Order 14110) directed the Secretary of Homeland Security to establish an AI Safety and Security Board, tasked with providing "advice, information, or recommendations" to the Secretary and the critical infrastructure community regarding its use of AI.[3] As an initial step, DHS, in close consultation with the Board, has developed this voluntary Roles and Responsibilities Framework for AI in Critical Infrastructure (the "Framework") to recommend the roles and responsibilities of entities across the AI ecosystem for the safe, secure, and resilient development and deployment of AI to critical infrastructure.*

* This Framework expresses standards of practice and does not establish any legal right, privilege, or requirement.

# II. Scope

This Framework proposes a model of shared and separate responsibilities for the safe and secure use of AI in critical infrastructure. For this purpose, the Framework:

- Recommends risk- and use case-based mitigations to reduce the risk of harm to **critical infrastructure systems** and the people served by them when developing and deploying AI, as well as the potential for harms to cascade in a manner that could impact multiple sectors or create nationally significant disruptions if left unaddressed.
- Proposes a set of voluntary responsibilities across the roles of **cloud and compute infrastructure providers**, **AI model developers**, and **critical infrastructure owners and operators** in developing and deploying the AI-powered services upon which much of the country's critical infrastructure currently relies or will soon rely.
- Proposes a set of voluntary responsibilities for **civil society** and **the public sector** in advocating for those who use or are affected by these critical systems, supporting research to improve various aspects of new technologies, and advancing strong risk-management practices.
- Relies upon existing risk frameworks to enable entities to evaluate whether the use of AI for certain systems or applications carries severe risks that could result in harms to critical infrastructure assets, sectors, or other nationally significant systems that serve the American people. Further research on the relationships between these risk categories, as well as their mitigations, will help entities conduct this evaluation on a use-case basis.

Furthermore, this Framework complements and leverages information gathered from the AI and critical infrastructure security programs DHS coordinates, including the annual AI sector-specific risk assessment process for critical infrastructure established under Executive Order 14110 and the forthcoming National Infrastructure Risk Management Plan.

# III. Risks of AI to Critical Infrastructure

For the purposes of this Framework, we refer to safety and security risks in the context of their potential threats, hazards, vulnerabilities, and consequences to critical infrastructure and offer risk mitigations that can be implemented across the AI development and deployment lifecycle. Our assessment of these factors draws from previously published U.S. government memoranda and reports; in particular, the Safety and Security Guidelines for Critical Infrastructure, the National Security Memorandum on Critical Infrastructure Security and Resilience, and guidance on risks to civil rights and civil liberties from OMB Memorandum M-24-10 and the White House Blueprint for an AI Bill of Rights.[4]

DHS, through the Cybersecurity and Infrastructure Security Agency (CISA) and in coordination with other Sector Risk Management Agencies (SRMAs),[5] identified three categories of AI safety and security attack vectors and vulnerabilities across critical infrastructure: attacks using AI, attacks targeting AI systems, and design and implementation failures. For owners and operators of critical infrastructure whose essential services and functions the public depends on daily, understanding the nature of these vulnerabilities and addressing them accordingly is not merely an operational requirement but a national imperative.

The National Security Memorandum on Critical Infrastructure Security and Resilience (NSM 22) articulates an approach to categorizing risks to critical infrastructure based on the scale and severity of potential harms, which in turn enables the prioritization of risk management efforts. Below, we have mapped each of these categories to an explanation of specific risks related to critical infrastructure's use of AI. Risk categories assess adverse impacts and are listed in order of increasing scale:

1. **Asset-Level Risk**, including but not limited to disruption or physical damage to the operations, assets, or systems of critical infrastructure, or direct supplier resulting from high-risk use of AI. These risks can include design or deployment flaws that impair service to a population or locality.

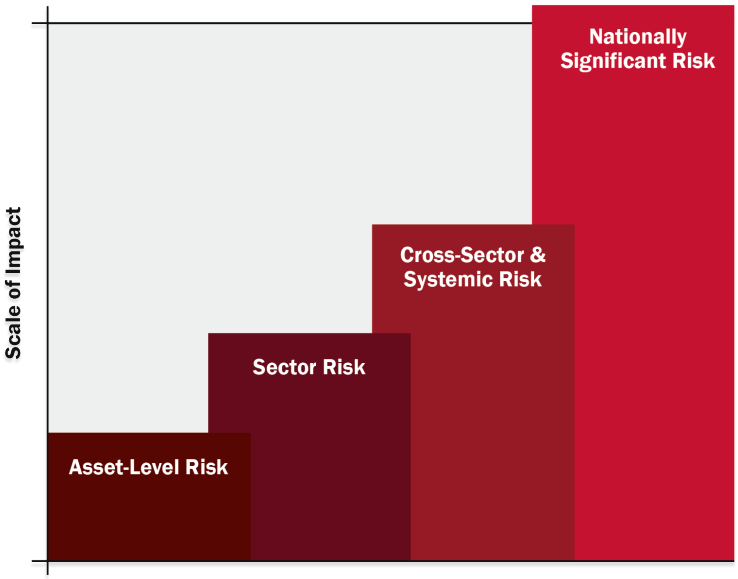


*Fig. 1. The Framework's risk categories align to the scale of potential harms.*

2. **Sector Risk**, including but not limited to risks to a set of assets impacting sector operations beyond an individual asset. These risks comprise, among others, operational failures of AI systems deployed in energy or water utilities and interruptions in the provision of vital services (such as medical services delivered by hospitals and financial services delivered by banks and other financial institutions), and the targeting of the election infrastructure subsector enabled by the misuse of AI.

3. **Systemic and Cross-Sector Risk**, including but not limited to disruptions to the information technology sector from AI-enabled cyberattacks and incidents that restrict access to critical services, negative environmental impacts associated with growing AI adoption, AI incidents resulting in significant financial loss, errors in AI-enabled sensing technologies that substantially hinder access to the services of critical infrastructure entities, disruptions in logistics supply chains due to failures in AI-enhanced processes, and other risks stemming from the increasingly interdependent nature of critical infrastructure.

4. **Nationally Significant Risk**, including but not limited to risks of AI causing widespread impact to safety or rights[6] or significantly assisting with the development, manufacture, or deployment of conventional, chemical, biological, radiological, or nuclear (CBRN) weapons.

This Framework suggests mitigations that, if implemented by the entities performing the relevant activities, can reduce the likelihood and severity of consequences associated with each risk category. Further, this framing of risks reveals the interdependent nature of these categories, where asset-level risks if left unaddressed can compound into sector-wide or cross-sector risks; conversely, mitigations designed to improve the safety or security of a critical asset may prevent or reduce the likelihood of a nationally significant consequence. This focus also acknowledges that the various choices made regarding how AI models are developed, how they can be accessed, and how they function within larger systems are critical to the impact they will have when deployed to broad segments of U.S. critical infrastructure. The public sector and civil society play a pivotal role in understanding and shaping this impact, so that benefits can be shared across sectors and harms can be prevented, mitigated, and, as necessary, remediated.

# IV. Roles and Responsibilities for the Safe and Secure Deployment of AI in Critical Infrastructure

## A. Overview

AI models will often be designed, implemented, and consumed by different entities in different contexts; therefore, responsibilities for managing AI safety and security risks will necessarily be distributed across multiple organizations. This Framework specifically evaluates the shared and separate responsibilities of cloud and compute infrastructure providers, AI model developers, critical infrastructure owners and operators, civil society, and the public sector across five foundational categories: securing environments, driving responsible model and system design, implementing data governance, ensuring safe and secure deployment, and monitoring performance and impact.

Interdependencies among these aforementioned entities must be addressed through concerted efforts to be transparent, using communication channels and data sharing to promote trust and common understanding. For example, critical infrastructure entities will be able to better assess whether models are safe and secure if AI developers provide customers with information about how relevant safety and security risks have been addressed through model design and testing.[7] Similarly, model developers and service providers can more efficiently assess risks to their IT environments and make informed procurement decisions if their cloud and compute infrastructure providers share documentation of hardware and software components and source vendors.[8] Furthermore, critical infrastructure entities can play a central role in improving and informing the AI development process by providing transparency into their deployments of AI, including the associated context and process of such deployments. Transparency serves an important purpose in connecting the entities identified in this Framework, enabling each to fulfill its respective responsibilities for safety and security.

Additionally, entities may play more than one role associated with the AI lifecycle. Recent developments in AI are increasingly enabling the configurability and tooling of models, closing the distance between technology providers and customers. A company that builds AI applications may also own the software it relies upon. A critical infrastructure entity may fine-tune an off-the-shelf AI model for certain bespoke services. The increasing overlap between services and capabilities across the AI supply chain has also caused uncertainty regarding where and with whom the core responsibilities for AI safety and security reside.[9] This Framework provides recommendations to help each entity assess its own obligations for AI safety and security while encouraging all entities to cooperate and communicate with one another to ensure the fulfillment of shared responsibilities.

## B. Key Terms and Structure

Entities across the AI ecosystem have shared and distinct responsibilities for the safe and secure development and deployment of AI in critical infrastructure. This Framework defines roles and responsibilities as follows:

- **Entity** may refer to a cloud and compute infrastructure provider, AI model developer, critical infrastructure owner and operator, civil society organization, or public sector government.
- **Roles** are defined broadly, encompassing a core service that an entity, or a group of entities,

provides to support the development and deployment of AI for a variety of uses, including critical infrastructure use cases. They include cloud and compute infrastructure providers, AI developers, critical infrastructure owners and operators, civil society, and the public sector. Each role may encompass multiple types of entities (e.g., AI developers include model developers as well as application developers), but these entities are grouped together based on the similarity of their responsibilities for AI safety and security in critical infrastructure.

- **Responsibilities** include technical risk mitigations, accountability and transparency mechanisms, rights-related protections, and testing benchmarks. As described in this Framework, responsibilities are akin to tasks performed by an entity and are grouped into the following shared goals: to secure environments, drive responsible model and system design, implement data governance, ensure safe and secure deployment, and monitor performance and impact.
- Entities should evaluate which responsibilities within this Framework are relevant to their overall role, as well as their specific activities related to the safe and secure development and deployment of AI in critical infrastructure.
- Entities should adopt and incorporate these responsibilities into their current or developing AI governance programs, including those embedded within technology development, procurement, and compliance processes. This Framework does not supersede existing legal or regulatory requirements to which entities are otherwise subject.

## C. The Framework

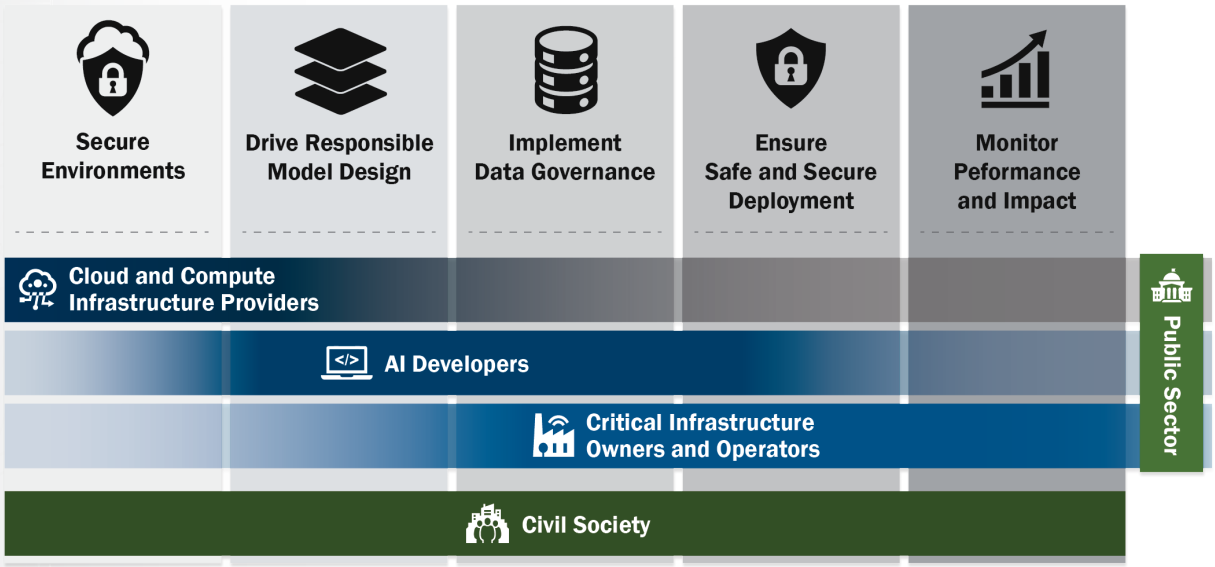


*Fig. 2. This Framework recommends best practices within a model of shared and separate responsibilities. Cloud and compute infrastructure providers, AI developers, and critical infrastructure owners and operators have responsibilities in each of the five substantive areas of AI safety and security at the top of this figure, but their primary responsibilities are indicated by their label's position in this graphic. The Framework recommends the public sector is responsible for ensuring that relevant private sector entities across all sectors are appropriately protecting the rights of individuals and communities, and civil society helps to advance research and individual rights across all substantive AI safety and security areas.*

# I. Cloud and Compute Infrastructure Providers

Cloud and compute infrastructure providers enable access to on-demand, scalable computing resources required to build, tune, and run AI models. Customers may also procure infrastructure from these entities to host model development and deployment on their premises. While these entities' responsibilities cluster in several categories, this Framework focuses on recommendations for how they can provide reliable, resilient, and secure-by-design AI products and services to critical infrastructure by ensuring data integrity, availability, and confidentiality.

| Cloud and Compute Infrastructure Providers Responsibilities Overview | | | | |
|---|---|---|---|---|
| **A. Secure Environments** | **B. Drive Responsible Model and System Design** | **C. Implement Data Governance** | **D. Ensure Safe and Secure Deployment** | **E. Monitor Performance and Impact** |
| 1. Vet hardware and software suppliers<br>2. Institute best practices for access management<br>3. Establish vulnerability management<br>4. Manage physical security | 1. Report vulnerabilities | 1. Keep data confidential<br>2. Ensure data availability | 1. Conduct systems testing | 1. Monitor for anomalous activity<br>2. Prepare for incidents<br>3. Establish clear pathways to report harmful activities |

## A. Secure Environments

1. Vet hardware and software suppliers: Cloud and compute infrastructure providers should review hardware and software in the supply chain to ensure the reliability and security of their components.[10] Supply chain risk management practices have been set out in a hardware bill of materials framework, a software bill of materials framework, and a software acquisition handbook that provide guidance on what information is appropriate to collect, depending on the circumstances.[11] Cloud and compute infrastructure providers should adopt these or other generally accepted frameworks to vet hardware and software suppliers.

2. Institute best practices for access management: Cloud and compute infrastructure providers should enable or implement best practices for monitoring and managing access to systems, models, or data sources by all users, devices, and applications.

3. Establish vulnerability management: Cloud and compute infrastructure providers should use vulnerability management methods to scan, or enable customers to scan infrastructure for threats, including threats stemming from AI. Providers should conduct or enable software security reviews, penetration testing, and audits by external parties, and they should build resilience to potential supply chain attacks in critical infrastructure applications.[12]

4. Manage physical security: Cloud and compute infrastructure providers should establish a layered physical security model. While circumstances may vary, a successful physical security model will generally include a) perimeter fencing and vehicle barriers, b) electronic access cards, c) access logs, d) 24/7 activity monitoring, and e) server and hardware hardening against tampering.[13]

## B. Drive Responsible Model and System Design

1. Report vulnerabilities: Where relevant, cloud and compute infrastructure providers should follow standard, coordinated vulnerability processes when reporting vulnerabilities that could affect model and system design processes.

## C. Implement Data Governance

1. Keep data confidential: Cloud and compute infrastructure providers should reduce the risk that personal or confidential customer data used to train or fine-tune models is exposed, leaked, or attacked by encrypting or enabling encryption methods for data at rest and in transit. Methods for doing so include making encryption available to customers and employing other best practices for data protection.

2. Ensure data availability: Cloud and compute infrastructure providers should utilize high-availability networking (i.e., networks that consistently operate at an optimal level without manual intervention) and backup plans in close cooperation with customers to ensure resiliency in the context of critical services.

## D. Ensure Safe and Secure Deployment

1. Conduct systems testing: Cloud and compute infrastructure providers should test the system environment to ensure the continued availability of services in a range of outage scenarios. This

security testing may be undertaken either in-house or through a reliable third-party provider. Where possible, cloud compute infrastructure providers should provide tools and diagnostics to customers to facilitate operational safety and security testing of their compute environment.

## E. Monitor Performance and Impact

1. Monitor for anomalous activity: Cloud and compute infrastructure providers should use appropriate tools to analyze or enable their customers to analyze network activity for potential threats and abuse.

2. Prepare for incidents: Cloud and compute infrastructure entities should work with providers to plan escalation, investigation, recovery, and communication processes in the event of an incident that involves unauthorized access to systems or data. If such unauthorized access occurs because of a cybersecurity, physical-security, insider-threat or other incident, entities should execute such plans.

3. Establish clear pathways to report harmful activities: Cloud providers should work with their customers to establish clear pathways for reporting suspicious or harmful activity, adhere to state, local, and federal reporting requirements, and work with public- and private-sector researchers to mitigate associated harms. Where relevant, entities should make use of existing incident reporting channels, such as Information Sharing and Analysis Centers (ISACs).

# II. AI Developers

AI developers are defined here as entities that develop, train, and/or enable access to AI models or applications, including through their own or third-party platform services and tools. They may develop the models themselves, modify or enable access to third-party models, provide software tools that enable developers to use or configure AI models, and/or deploy models into downstream applications for customers. These developers have safety and security responsibilities across the AI lifecycle for critical infrastructure, particularly as they relate to driving responsible model and application design, testing, and maintenance over time. Some of the safety and security responsibilities discussed below are shared by all types of AI developers, whereas other responsibilities apply specifically to certain types of AI developers, depending on their access to the upstream AI model or the downstream AI application, or whether the model's weights are widely available. This Framework aggregates these entities—developers of AI models, platforms, and applications—into a single category because they play a similar role in developing AI technologies that will ultimately be used by critical infrastructure entities.

For responsibilities in this section related to preventing dual-use models from being deliberately misused to carry out attacks on critical infrastructure, AI model developers are encouraged to refer to the AI Safety Institute's draft guidelines, *Managing Misuse Risk for Dual-Use Foundation Models*, for additional and more detailed recommendations. [14]

**AI Developers Responsibilities Overview**

| A. Secure Environments | B. Drive Responsible Model and System Design | C. Implement Data Governance | D. Ensure Safe and Secure Deployment | E. Monitor Performance and Impact |
|---|---|---|---|---|
| 1. Manage access to models and data | 1. Incorporate Secure by Design principles | 1. Respect individual choice and privacy | 1. Use a risk-based approach when managing access to models | 1. Monitor AI models for unusual or adversarial activity |
| 2. Prepare incident response plans | 2. Evaluate dangerous capabilities of models | 2. Promote data and output quality | 2. Distinguish AI-generated content | 2. Identify, communicate, and address risks |
| | 3. Ensure alignment with human-centric values | | 3. Validate AI system use | 3. Support independent assessments |
| | | | 4. Provide meaningful transparency to customers and the public | |
| | | | 5. Evaluate real-world risks and possible outcomes | |
| | | | 6. Maintain processes for vuln. reporting and mitigation | |

## A. Secure Environments

1. Manage access to models and data: AI developers should help ensure that components of an AI system that play a crucial role in its overall security posture, such as model weights, training data, and source code, are protected from unauthorized access.** Where applicable, developers should implement controls to detect and prevent attempts to access, modify, and exfiltrate confidential information.[15]

2. Prepare incident response plans: AI developers should create clear reporting and evaluation processes to respond quickly to incidents, including insider threats. Plans should be regularly reviewed and incorporated into training and tabletop exercises, which prepare employees to efficiently assess and identify risks, follow escalation pathways, and contain the impact of an incident. Lessons learned from incidents should be incorporated into plans to improve responses to future events.[16]

## B. Drive Responsible Model and System Design

1. Incorporate Secure by Design principles: AI developers should ensure their products are developed from the outset with an emphasis on security, using established frameworks, such as CISA's Secure by Design Pledge, to guide and publicly communicate their approach to building secure AI. Developers should provide regular updates to public-facing security policies to reflect progress towards security goals, train their workforce on how to identify and mitigate relevant security vulnerabilities, provide current information for customers to consider, and incorporate new research and guidance on security best practices.[17]

2. Evaluate dangerous capabilities of models: AI model developers should establish and adhere to a strategy to identify capabilities associated with autonomous activity, physical and life sciences, cybersecurity, and other capabilities that could impact critical infrastructure when deployed in relevant high-risk contexts.

3. Ensure alignment with human-centric values: AI model developers should ensure, to the best of their ability, that AI models reflect human values and goals, with the ultimate objective of ensuring they are helpful, accurate, unbiased, and transparent.[18] AI application developers should align use cases with values that respect civil rights, civil liberties, and applicable laws in partnership with relevant civil society.[19]

## C. Implement Data Governance

1. Respect individual choice and privacy: AI developers should ensure that effective data management for AI systems considers an individual's legal rights, stated choices, and reasonable expectations of privacy. AI developers should implement privacy best practices, including data minimization, and comply with applicable regulations when collecting, processing, retaining, and transferring personal information.[20]

2. Promote data and output quality: Given the number and variability of data sources used as inputs to train AI models, AI developers should consistently evaluate the quality of model inputs and use methods that enhance the quality and reliability of outputs, such as data filtering, fine-tuning, and

** This responsibility does not take a position on securing these components if they are made widely accessible to members of the public for use and research.

classification, to help prevent unintended or harmful outcomes of the model.

## D. Ensure Safe and Secure Deployment

1. Use a risk-based approach when managing access to models: AI model developers should perform a risk assessment prior to making model weights widely available. The assessment should consider the benefits of openness, including the impact to AI safety and security research, against the risks of misuse.[21]

2. Distinguish AI-generated content: Where technically feasible and commercially reasonable, AI developers are encouraged to ensure that AI-generated or manipulated content, such as code, text, images, audio, or video, can be clearly identified at the time and point of origin, and therefore distinguishable from human-generated content, consistent with legal requirements and safety best practices. AI developers should continue to update their evaluation methods as research in these areas progresses.

3. Validate AI system use: AI developers should test for general reliability and robustness, using available benchmarks where possible, to help ensure that the AI system will act as planned under normal conditions and a wide range of possible conditions.[22] Where benchmarks do not yet exist, AI developers should support internal and external research into relevant testing, evaluation, validation, and verification (TEVV) approaches, including application-specific benchmarks.

4. Provide meaningful transparency to customers and the public: AI developers should provide information that enables their critical infrastructure customers to conduct their own risk assessments and make informed decisions about when and how to use AI. Information provided by AI developers may include information about training data and model architecture, performance on benchmarks of interest and results of evaluations to detect dangerous capabilities, and details of risk management practices, including safety and security measures.[23]

5. Evaluate real-world risks and possible outcomes: AI developers should subject models and applications to evaluations that test for possible biases, failure modes, or vulnerabilities that may result in harmful outcomes when they are intended or designed to be deployed in high-risk contexts. Model developers should implement risk management policies that include testing procedures, such as AI red-teaming,[24] and risk thresholds that can be used to restrict further development or deployment of models if corresponding safety measures are not effective in reducing risks below such thresholds.[25] AI application developers should assess performance for intended use cases and carefully consider whether the AI application could create inadvertent disparate impacts.

6. Maintain processes for vulnerability reporting and mitigation: Consistent with the public interest and strong security protocols, AI developers should conduct timely remediation activities for known vulnerabilities that could result in harms to critical infrastructure. Where practicable, entities should consider disclosing their remediation activities to customers, partners, and regulators through ISACs or other appropriate channels.

## E. Monitor Performance and Impact

1. Monitor AI models for unusual or adversarial activity: AI developers should monitor and/or enable customers to monitor indicators of compromise[26] that may suggest unusual behavior or malicious activity posing risk to critical infrastructure, including AI-enabled cyber-attacks or adversarial

manipulation, model drift, or data poisoning. In certain high-trust deployments, it may be appropriate for developers to assist customers in carrying out these functions themselves, owing to the nature of the data, privacy, or security concerns.

2. Identify, communicate, and address risks: AI developers should establish processes for documenting and communicating newly identified model risks and for prioritizing risks based on impact, likelihood, and available resources and mitigations. AI developers should establish measures such as bug bounties, regular cadences for the reporting and patching of vulnerabilities, and processes for receiving and analyzing customer reporting.

3. Support independent assessments: AI model developers should enable qualified and trusted third parties to evaluate models that present Nationally Significant Risks, such as CBRN-related capabilities[27] and/or heightened risks to critical infrastructure entities and their consumers.

# III. Critical Infrastructure Owners and Operators

Critical infrastructure owners and operators manage the secure operation and maintenance of critical systems, which increasingly rely on AI to reduce costs, improve reliability, and boost efficiency. These critical infrastructure entities typically interact directly with AI applications or platforms that enable them to configure AI models for specific use cases. While AI use cases vary broadly across sectors, both in terms of their functions and risks, the ways in which AI models and systems are deployed have important safety and security implications for critical services, as well as the individuals who consume such services.[28]

**Critical Infrastructure Owners and Operators Responsibilities Overview**

| A. Secure Environments | B. Drive Responsible Model and System Design | C. Implement Data Governance | D. Ensure Safe and Secure Deployment | E. Monitor Performance and Impact |
|---|---|---|---|---|
| 1. Secure existing IT infrastructure | 1. Use responsible procurement guidelines<br>2. Evaluate AI use cases and associated risks<br>3. Implement safety mechanisms<br>4. Establish appropriate human oversight | 1. Protect customer data used to configure or fine-tune models<br>2. Manage data collection and use | 1. Maintain cyber hygiene<br>2. Provide transparency and consumer rights<br>3. Build a culture of safety, security, and accountability for AI<br>4. Train the workforce | 1. Account for AI in incident response plans<br>2. Track and share performance data<br>3. Conduct periodic and incident-related TEVV<br>4. Measure impact<br>5. Ensure system redundancy |

## A. Secure Environments

1. Secure existing IT infrastructure: Critical infrastructure entities should apply internationally accepted standards and practices[29] where applicable and where AI systems will be deployed, including by managing deployment-environment governance, ensuring robust architecture, hardening configurations, and protecting networks from threats.[30]

## B. Drive Responsible Model and System Design

1. Use responsible procurement guidelines: Critical infrastructure entities should confirm with AI developers that their AI products and services are tested, evaluated, validated, and verified by operational and subject matter experts. Entities should practice a "secure by demand" approach and evaluate both enterprise and product security to ensure they meet standards for cybersecurity, privacy, and data integrity (including data accuracy and consistency).[31]

2. Evaluate AI use cases and associated risks: When considering the use of an AI application or its underlying model, critical infrastructure entities should evaluate how AI may impact system operations prior to implementation, which includes a) the intended purpose of the AI application and expected use cases, b) possible failure modes associated with the system that uses or is reliant upon the AI application, and c) implications and mitigations related to issues of bias and fairness relevant to critical infrastructure.

3. Implement safety mechanisms: Critical infrastructure entities should implement controls in AI systems to prevent or mitigate the potential severity of safety-impacting outcomes that may be caused by automated decision-making processes that pose severe risks to critical infrastructure assets or employees.

4. Establish appropriate human oversight: Critical infrastructure entities should incorporate appropriate human involvement for making or informing consequential decisions that could negatively impact critical infrastructure assets, services, or consumers. For such decisions, entities should establish meaningful human oversight and specify the extent to which owners and operators should rely on AI-generated outputs, predictions, and/or forecasts.

## C. Implement Data Governance

1. Protect customer data used to configure or fine-tune models: To mitigate risks associated with the use of proprietary or private data, critical infrastructure entities should protect customer data from improper exposure, especially when such data is used when training or fine-tuning models. Entities should collect, process, retain, and transfer only as much customer data as is necessary to serve the specific task at hand, consistent with relevant customer authorizations and consents.

2. Manage data collection and use: Critical infrastructure entities should track and protect customer data used to fine-tune AI models or configure AI applications for their use cases. They should work with AI developers as needed to verify the integrity of external datasets purchased or procured for model training (including data accuracy, consistency, and security) to assess their suitability for use in the given context, where practicable.[32]

## D. Ensure Safe and Secure Deployment

1. Maintain cyber hygiene: Critical infrastructure entities should implement strong cybersecurity practices, such as those outlined in CISA's Cyber Performance Goals, to maintain controls for AI systems.[33]

2. Provide transparency and consumer rights: Critical infrastructure entities should provide meaningful transparency regarding their use of AI to provide goods, services, or benefits to the public. Entities should work with their stakeholders, including local governments, communities, and members of the public, to determine the appropriate types, level, and cadence of information disclosures, while ensuring the security of intellectual property. Where practicable, entities that use AI systems that may adversely and materially affect individuals or communities should give impacted individuals an opportunity to obtain an explanation of how the system affects them.

3. Build a culture of safety, security, and accountability for AI: For the use of AI for critical systems, critical infrastructure owners and operators should ensure that their executive leadership is engaged in key decisions, which are supported by appropriate governance policies and procedures.

4. Train the workforce: Critical infrastructure entities should train their workforce on appropriate uses of AI and security vulnerabilities that specifically impact employees of critical infrastructure owners and operators, such as phishing attacks, malware, vulnerable devices, poor password hygiene, or data poisoning.

## E. Monitor Performance and Impact

1. Account for AI in incident response plans: If an incident impacting critical infrastructure occurs, critical infrastructure entities should be prepared with an up-to-date general incident response plan. The plan should instruct how to cease the operation of an AI system, commence operation of a backup system to provide continued availability of the critical infrastructure service, notify the appropriate government authority, inform impacted customers and other stakeholders as appropriate, and safely withdraw access to the AI system. Entities should develop plans and protocols to roll back AI systems if an issue arises.

2. Track and share performance data: Where applicable, critical infrastructure entities should work with model or system developers to define processes for sharing information and/or data on how an AI system performed. Critical infrastructure entities should also consider sharing the results of any evaluation of that performance data to help the developer better understand the relationship between model behavior and real-world outcomes. This data sharing should be conducted in a manner that protects personal data.

3. Conduct periodic and incident-related testing, evaluation, validation, and verification: As part of a continuous system monitoring/risk management process, critical infrastructure entities should periodically assess the sufficiency and efficacy of tests and measurements for implementing repairs or upgrades.[34]

4. Measure impact: Critical infrastructure entities should measure the impact that AI has on the overall system into which the model is integrated on an ongoing basis, including disparate impact to individuals or communities who consume or are otherwise affected by the system.

5. Ensure system redundancy: Critical infrastructure entities should incorporate redundancy to minimize the impacts of disruptions caused by natural disasters, accidents, or other events.

# IV. Civil Society

Civil society—those organizations distinct from industry and government—includes non-governmental organizations, labor unions, charitable organizations, professional associations, foundations, academia, research institutions, and others. The roles of civil society organizations are diverse, from non-governmental organizations that engage with private-sector AI companies to research institutions that contribute to a culture of innovation. As a sector, civil society contributes to standards, frameworks, and solutions that can help to measure and communicate the impact of technology on individuals and communities, in partnership with government and industry. Civil society plays a key role in supporting, standardizing, and improving safety, privacy, fairness, and security mitigations across the entire AI lifecycle, protecting the public and fostering trustworthiness.

| Civil Society Responsibilities Overview | | | | |
|---|---|---|---|---|
| **A. Secure Environments** | **B. Drive Responsible Model and System Design** | **C. Implement Data Governance** | **D. Ensure Safe and Secure Deployment** | **E. Monitor Performance and Impact** |
| 1. Actively engage in developing and communicating standards, best practices, and metrics alongside government and industry<br>2. Educate policymakers and the public<br>3. Inform guiding values for AI system development and deployment<br>4. Support the use of privacy-enhancing technologies<br>5. Consider critical infrastructure use cases for red-teaming standards<br>6. Continue to drive and support research and innovation | | | | |

1. Actively engage in developing and communicating standards, best practices, and metrics alongside government and industry: Civil society should support the development and communication of AI safety standards for the use of AI by critical infrastructure and concrete metrics to measure the impact of sector-specific applications. These standards and metrics must be rights-affirming, with tools to define, evaluate, and monitor against bias at the development stage.

2. Educate policymakers and the public: Civil society entities should develop AI educational resources or identify appropriate contributions to help inform policymakers and the public about the uses, benefits, and risks of AI.

3. Inform guiding values and safeguards for AI system development and deployment: Civil society should work with the public and with government to formulate—and with cloud and compute infrastructure providers and AI developers to implement—guiding values and safeguards, including appropriate rules, policies, and procedures, to develop and deploy AI systems that are transparent and that protect privacy, civil rights, human rights, and societal well-being.[35]

4. Support the use of privacy-enhancing technologies: Civil society should work with industry where

applicable to identify opportunities to drive adoption of privacy enhancing technologies (PETs) for collecting, processing, and training data used for AI.

5. Consider critical infrastructure use cases for red-teaming standards: Civil society should engage with AI developers where applicable to develop practical red-teaming standards that can be readily adopted by critical infrastructure across a variety of use cases and risk thresholds.

6. Continue to drive and support research and innovation: Civil society should advance fundamental research in AI applications that supports the safe development and deployment of AI in critical infrastructure and key sociotechnical concepts in AI. Such research should focus on equality, equity, and democratic values and aim to foster responsible innovation.

# V. Public Sector

The public sector—including government agencies from the federal, state, local, tribal, and territorial entities—serves and protects the American people and its institutions. The public sector has a responsibility to ensure that relevant private sector entities across all sectors are appropriately protecting the rights of individuals and communities, as well as a responsibility to respond and support the American public in times of crisis or emergency.*** In light of these important duties, this Framework recommends the following practices for public sector entities in the context of their own use of AI[36], as well as their efforts to promote the safe and secure use of AI in critical infrastructure.

| | Public Sector Roles and Responsibilities |
|---|---|
| **Cloud and Compute Infrastructure Providers** | 1. Deliver essential services and emergency response<br>2. Drive global AI norms<br>3. Responsibly leverage AI to improve the functioning of critical infrastructure<br>4. Advance standards of practice through law and regulation<br>5. Engage community leaders<br>6. Enable foundational research into AI safety and security<br>7. Support critical infrastructure's safe and secure adoption of AI<br>8. Develop oversight |
| **AI Developers** | |
| **Critical Infrastructure Owners and Operators** | |

1. Deliver essential services and emergency response: The public sector should ensure that its use of AI supports and never conflicts with core governmental functions, namely essential services, life and safety, emergency response, and economic and community support.

2. Drive global AI norms: The United States is the world leader in AI and will work with its partners to lead in the establishment of strong norms and standards around AI safety and security. The federal government should engage with international partners on AI to identify common threats, drive international regulations and standards, and converge around shared responsibilities to protect all global citizens.

*** For example, local governments are most often multi-sector by default and risks to infrastructure and/or operations can pose an immediate risk to the life and safety of citizens. To wit, a major city will often have membership in the Emergency, Energy, Government/Multi-State, Technology, Transportation, Health, and Water/Wastewater Information Sharing and Analysis Centers.

3. Responsibly leverage AI to improve the functioning of critical infrastructure: The public sector should improve efficiencies, increase affordability and availability of critical services, and lead by example in transparency and communication to the public. It should prioritize the development of, and funding for, programs that advance responsible AI practices in government services. Public sector entities should engage with civil society and each other regarding the public sector's use of AI and avoid using AI in a manner that produces discriminatory outcomes, infringes upon personal privacy, or violates other legal rights. Public sector entities should not fund discriminatory technologies.

4. Advance standards of practice through law and regulation: The federal government has the opportunity to advance standards of practice through statutory and regulatory action. In doing so, the government must ensure that it does not stifle innovation, especially given the dynamic and rapidly evolving landscape of AI. Laws and regulations should protect individuals' fundamental rights, help drive innovation, advance the harmonization of different legal requirements, simplify compliance, and clarify incident reporting processes.

5. Engage community leaders: The public sector should work with community leaders to anticipate and understand the impact of AI on constituents, workforces, and institutions with a special focus on vulnerable populations. The federal government should coordinate with local governments to ensure impacts are measured, understood, and used to inform federal policy, where relevant.

6. Enable foundational research into AI safety and security: The public sector should work with AI developers and researchers to test, analyze, and monitor how technical advances in AI may impact the safety and security of critical services, such as financial, educational, healthcare, and other services. Public sector leaders should partner with academic institutions and utilize AI funding initiatives like the National AI Research Resource (NAIRR) to lead and participate in efforts to build standards around strong AI safety and security practices across industries and sectors.

7. Support critical infrastructure's safe and secure adoption of AI: The public sector should articulate support for the responsible use of AI by critical infrastructure where such use will be beneficial and identify circumstances in which the use of AI would be inappropriate.

8. Develop oversight: DHS and the Board should assess the continued relevance of this Framework and make public its oversight and assessment mechanisms.

# V. Conclusion

Recent advances in AI present extraordinary possibilities to improve the functioning of critical infrastructure if associated risks can be effectively managed. This Framework provides a foundation for how leaders across sectors, industries, and governments can help advance this field by assuming and fulfilling shared and separate responsibilities for AI safety and security, both within their organizations and as part of their interactions with others. This Framework will succeed if, among other achievements, it further strengthens harmonization of AI safety and security practices, improves the delivery of critical services enabled by AI, enhances trust and transparency across the AI ecosystem, advances research into safe and secure AI for critical infrastructure, and ensures that civil rights and civil liberties are protected by all entities.

The extraordinary scale and impact of U.S. critical infrastructure underscores the importance of aligning on an approach to developing and deploying AI for critical systems, consistent with our nation's core values. This Framework seeks to complement and advance the AI safety and security best practices established by the White House Voluntary Commitments and Blueprint for an AI Bill of Rights, the OMB M-24-10 Memorandum on AI, the work of the AI Safety Institute, and the DHS Safety and Security Guidelines for Critical Infrastructure Owners and Operators, among others.

## A. Desired Outcomes of Framework

1. **Strengthened harmonization of foundational safety and security practices.** Recent advances in AI capabilities have led to a proliferation of AI safety and security guidance, proposed standards, and principles. But unless entities can agree on how to operationalize these resources, they risk becoming obsolete. This Framework draws upon established guidance, as well as the insights of stakeholders from across the AI ecosystem, to offer a model of shared and separate responsibilities, designed to be adopted, implemented, and updated to keep pace with changing technologies. Specifically, this Framework should be used alongside other government-issued guidance to inform corporate policies, agreements between entities, technical standards for AI development, and statutory or regulatory action that will ultimately govern important aspects of how AI is developed and deployed to U.S. critical infrastructure. Further, the wide adoption of this Framework by U.S. entities will help to shape international norms and standards, given the global influence of American leadership and innovation. This harmonization, across the country and around the world, is vital to the driving force of innovation that will define the power of AI to strengthen and build resilience in critical infrastructure for years to come.

2. **Improved planning and delivery of critical services.** Operating and ensuring the safety and security of a critical system, such as an oil pipeline or air traffic system, requires the careful orchestration of multiple processes, including advanced planning, real-time monitoring, and sophisticated forecasting, among others. Each of these processes stands to benefit greatly from AI, both in terms of their individual execution and their coordination together. While the owners and operators of these systems must always carefully consider the risks of implementing nascent technologies into vital systems, critical infrastructure entities have historically been leaders in researching, piloting, and deploying new technologies to serve the American public. Today, recent advances in AI and generative AI present new opportunities for leadership and innovation, while also making clear the need to assess risks and build in critical protections for consumers. By implementing the relevant components of this Framework into the policies and processes that govern the way critical infrastructure builds or procures its AI systems, owners and operators can help shape the growing field of AI safety and security and exemplify and further the positive impact

that AI can have on society.

3. **Enhanced trust and transparency across the AI ecosystem.** Trust and transparency work hand-in-hand in any complex ecosystem, and AI is no exception. For technology providers, including cloud and compute infrastructure providers and AI developers, providing transparency is critical to building customer trust and confidence; furthermore, the information they share should be meaningful to the intended audience and calibrated to the applicable level of risk. For certain high-risk contexts, technology providers should additionally be able to provide assurances that an AI system operates in a reliable manner. Strong transparency practices among entities enable downstream deployers, such as critical infrastructure entities, to address and assuage concerns from those consumers who interact directly or indirectly with AI in critical systems. Accordingly, this Framework views trust and transparency as shared responsibilities that take different forms, depending on the role of an entity and their position along the AI ecosystem. By adopting these shared and separate responsibilities, entities can help drive key procurement decisions, enable meaningful information exchanges, support trusted third-party involvement, and foster collaboration and common understanding across the ecosystem.

4. **Advancement of research into AI safety and security outcomes for critical infrastructure.** Today's AI research community is rapidly expanding to encompass and keep pace with the development and testing of novel capabilities and risks of increasingly advanced AI models. Further research and development should help critical infrastructure operationalize and deploy AI use cases, including those that do not leverage the most advanced AI models, and manage risks that emerge from the application of AI in specific contexts. This Framework calls for all entities to support and drive research focused on the use of AI for critical services, from energy management to healthcare, to foster a better understanding of AI's potential use cases in these sectors, as well as the relationship between AI model architecture and real-world outcomes. Furthermore, such research can help inform efforts to develop effective and risk-based AI regulation, which relies on a concrete understanding of benefits, harms, and mitigations.

5. **Respect for civil rights.** The use of AI in critical infrastructure and its corresponding costs and benefits will vary depending on the specific application, the context of the sector and use case, and many other factors. Nevertheless, the consideration of privacy, civil rights, and civil liberties is foundational and must be carried across all AI systems. Accordingly, this Framework makes safeguarding civil rights, identifying disparate impacts, and mitigating harm shared responsibilities across the full AI ecosystem that supports the development and deployment of AI in critical infrastructure.

# Appendix A: AI Roles and Responsibilities Matrix

| | Secure Environments | Drive Responsible Model and System Design | Implement Data Governance | Ensure Safe and Secure Deployment | Monitor Performance and Impact | Public Sector |
|---|---|---|---|---|---|---|
| **Cloud and Compute Infrastructure Providers** | • Vet hardware and software suppliers<br>• Institute best practices for access management<br>• Establish vulnerability management<br>• Manage physical security | • Report vulnerabilities | • Keep data confidential<br>• Ensure data availability | • Conduct systems testing | • Monitor for anomalous activity<br>• Prepare for incidents<br>• Establish clear pathways to report harmful activities | • Deliver essential services and emergency response<br>• Drive global AI norms<br>• Responsibly leverage AI to improve the functioning of critical infrastructure<br>• Advance standards of practice through law and regulation<br>• Engage communicty leaders<br>• Enable foundational research into AI safety and security<br>• Support critical infrastructure's safe and secure adoption of AI<br>• Develop oversight |
| **AI Developers** | • Manage access to models and data<br>• Prepare incident response plans | • Incorporate Secure by Design principles<br>• Evaluate dangerous capabilities of models<br>• Ensure alignment with human-centric values | • Respect individual choice and privacy<br>• Promote data and output quality | • Use a risk-based approach when managing access to models<br>• Distinguish AI-generated content<br>• Validate AI system use<br>• Provide meaningful transparency to customers and the public<br>• Evaluate real-world risks and possible outcomes<br>• Maintain processes for vuln. reporting and mitigation | • Monitor AI models for unusual or adversarial activity<br>• Identify, communicate, and address risks<br>• Support independent assessments | |
| **Critical Infrastructure Owners and Operators** | • Secure existing IT infrastructure | • Use responsible procurement guidelines<br>• Evaluate AI use cases and associated risks<br>• Implement safety mechanisms<br>• Establish appropriate human oversight | • Protect customer data used to configure or fine-tune models<br>• Manage data collection and use | • Maintain cyber hygiene<br>• Provide transparency and consumer rights<br>• Build a culture of safety, security, and accountability for AI<br>• Train the workforce | • Account for AI in incident response plans<br>• Track and share performance data<br>• Conduct periodic and incident-related TEVV<br>• Measure impact<br>• Ensure system redundancy | |
| **Civil Society** | • Actively engage in developing and communicating standards, best practices, and metrics alongside government and industry<br>• Educate policymakers and the public<br>• Inform guiding values for AI system development and deployment<br>• Support the use of privacy-enhancing technologies<br>• Consider critical infrastructure use cases for red-teaming standards<br>• Continue to drive and support research and innovation | | | | | |

# Appendix B: Glossary

Many terms in this document have various definitions across laws and guidance. For clarity, we have included some helpful terms here but defer to entities to use the definitions most appropriate to their activities.

**ARTIFICIAL INTELLIGENCE (AI).** 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. *See* 15 U.S.C. § 9401(3).

**AI APPLICATION.** A computer program hosted on an information technology system that is designed to perform a specific set of AI related tasks or requirements. An AI application is created by a developer and can, but does not necessarily, involve the creation, modification, collection, processing, and/or presenting of data or algorithms. *See* NIST, *Common Platform Enumeration: Naming Specification Version* 2.3, at 3 (Aug. 2011), https://doi.org/10.6028/NIST.IR.7695.

**AI DEPLOYERS.** Entities that put an AI system into use under authority. They may use the AI system to make decisions that impact end-users or use an AI system produce AI models, applications, platforms, or services. They may enable access to AI, develop the models themselves, or create software tools that enable organizations to use and configure AI models for specific applications.

**AI MODEL.** 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. *See* Executive Order 14110, § 3(c).

**AI PLATFORM.** Entities that operate a device, operating system, or virtual environment on which AI models or systems can be developed, installed, modified, integrated, distributed, or run by an end user. An AI platform is a dynamic ecosystem that depends on multiple technologies, tenant developers and/or end users to reliably deliver its services. AI platforms are a subcategory of developers and may support a range of AI system services, including deployment or service delivery. See NIST, *Trustworthy Platforms*, https://www.nist.gov/trustworthy-platforms.

**AI RED-TEAMING.** 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. *See* Executive Order 14110, § 3(d).

**AI SYSTEM.** Any data system, software, hardware, application, tool, or utility that operates in whole or in part using AI. *See* Executive Order 14110, § 3(e).

**CIVIL SOCIETY.** Organizations, distinct from industry and government, that include, inter alia, non-governmental organizations, labor unions, charitable organizations, professional associations, foundations, academia, and research institutions.

**CLOUD AND COMPUTE INFRASTRUCTURE.** Entities that provide access to on-demand, scalable computing resources required to build, tune, and run AI models. Customers may also procure infrastructure from these entities in order to host model development and deployment on-premises.

**CRITICAL INFRASTRUCTURE.** The 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. *See* 42 U.S.C. § 5195c(e). Critical infrastructure entities manage the secure operation and maintenance of critical systems, some of which use (and some of which design) AI models.

**END USER.** The ultimate consumer of a finished product or service. *See* 22 U.S.C. § 8541(5).

**FOUNDATION MODEL.** An AI model that is trained on broad data; generally uses self-supervision; contains at least tens of billions of parameters; is applicable across a wide range of contexts; and that exhibits, or could be easily modified to exhibit, high levels of performance at tasks. *See* Executive Order 14110, § 3(k).

**MODEL CARD.** Short documents accompanying trained machine learning models that provide benchmarked evaluation in a variety of conditions, such as across different cultural, demographic, or phenotypic groups that are relevant to the intended application domains. Model cards disclose the context in which models are intended to be used, details of the performance evaluation procedures, and other relevant information. *See* Mitchell et al., *Model Cards for Model Reporting*, *supra* note 7.

**MODEL WEIGHT.** A numerical parameter within an AI model that helps determine the model's outputs in response to inputs. *See* Executive Order 14110, § 3(u).

**PUBLIC SECTOR.** Federal, state, local, tribal, and territorial governments, sub-divisions, and officials.

# Appendix C: Notes

1 Federal law defines critical infrastructure as the "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." 42 U.S.C. § 5195c(e).

2 *See generally* The White House, *Blueprint for an AI Bill of Rights* (Oct. 2022), https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf; Office of Management & Budget, *Memorandum for the Heads of Executive Departments and Agencies on Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence and Appendix I* (Mar. 28, 2024), https://www.whitehouse.gov/wp-content/uploads/2024/03/M-24-10-Advancing-Governance-Innovation-and-Risk-Management-for-Agency-Use-of-Artificial-Intelligence.pdf (OMB, M-24-10) (list of uses of AI that are presumed to be rights- or safety-impacting).

3 The White House, *Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*, § 4.3(a)(v) (Oct. 30, 2023), https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

4 *See* The White House, *National Security Memorandum on Critical Infrastructure Security and Resilience*, NSM-22, (Apr. 30, 2024), https://www.whitehouse.gov/briefing-room/presidential-actions/2024/04/30/national-security-memorandum-on-critical-infrastructure-security-and-resilience/; DHS, *Safety and Security Guidelines for Critical Infrastructure Owners and Operators* (Apr. 26, 2024), https://www.dhs.gov/publication/safety-and-security-guidelines-critical-infrastructure-owners-and-operators; OMB, M-24-10, *supra* note 2, § 2; The White House, *Blueprint for an AI Bill of Rights*, *supra* note 2.

5 The National Security Memorandum on Critical Infrastructure Security and Resilience (NSM-22) identifies 16 critical infrastructure sectors and associated Sector Risk Management Agencies (SRMAs). SRMAs serve as day-to-day Federal interfaces for their designated critical infrastructure sector and conduct sector specific risk-management and resilience activities. *See* The White House, NSM-22, *supra* note 4.

6 *See* OMB, M-24-10, *supra* note 2, § 5.

7 This documentation could take the form of a "model card" or "factsheet." *See* Margaret Mitchell et al., *Model Cards for Model Reporting, Conference on Fairness, Accountability, and Transparency* (Jan. 29-31, 2019), https://arxiv.org/pdf/1810.03993; John Richards et al., *A Methodology for Creating AI Factsheets*, IBM (June 28, 2020), https://arxiv.org/pdf/2006.13796.

8 This documentation could take the form of a "Hardware Bill of Materials." *See* CISA, *Hardware Bill of Materials (HBOM) Framework for Supply Chain Risk Management* (Sept. 25, 2023), https://www.cisa.gov/resources-tools/resources/hardware-bill-materials-hbom-framework-supply-chain-risk-management.

9 Documenting your organization's AI supply chain is an essential component of secure development. *See* CISA and NCSCUK, *Guidelines for Secure AI System Development*, at 12, (Dec. 13, 2023), https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development.

10 *See* John Pendleton et al., Cloud Reassurance: A Framework to Enhance Resilience and Trust, Carnegie Endowment for International Peace, (Jan. 2024), https://carnegieendowment.org/research/2024/01/cloud-reassurance-a-framework-to-enhance-resilience-and-trust.

11 *See* NIST, *Cybersecurity Supply Chain Risk Management Practices for Systems and Organizations* (May 2022), https://csrc.nist.gov/pubs/sp/800/161/r1/final; CISA, *supra* note 8; CISA, *Software Acquisition Guide for Government Enterprise Consumers: Software Assurance in the Cyber-Supply Chain Risk Management (C-SCRM) Lifecycle* (Aug. 1, 2024), https://www.cisa.gov/resources-tools/resources/software-acquisition-guide-government-enterprise-consumers-software-assurance-cyber-supply-chain. *See also* NIST, *Security and Privacy Controls for Information Systems and Organizations, SP* 800-53, *Revision* 5 (Sept. 2020), https://doi.org/10.6028/NIST.SP.800-53r5.

12 *See* CISA, *A Guide to Critical Infrastructure Security and Resilience* (Nov. 2019), https://www.cisa.gov/resources-tools/resources/guide-critical-infrastructure-security-and-resilience.

13 *See* Pendleton et al., *supra* note 10.

14 *See* NIST, *Managing Misuse Risk for Dual-Use Foundation Models, Initial Public Draft* (July 2024), https://doi.org/10.6028/NIST.AI.800-1.ipd.

15 *See* CISA and NCSCUK, *supra* note 9, at 14.

16 *See* NIST, *supra* note 14.

17 *See* CISA, Secure by Design Pledge (2023), https://www.cisa.gov/securebydesign/pledge.

18 For example, AI model and platform developers may achieve this alignment by using appropriately representative datasets, considering how the system may operate regarding different populations and demographics, and conducting pre-deployment testing to identify and mitigate biases and other confounding variables. For normative and technical definitions of alignment, see NIST, *The Language of Trustworthy AI: An In-Depth Glossary of Terms* (Mar. 29, 2023), https://www.nist.gov/publications/language-trustworthy-ai-depth-glossary-terms.

19 Existing legal authorities—civil rights, non-discrimination, fair competition, consumer protection, and other vitally important legal protections—apply to the design and use of AI models and systems just as they apply to other practices. *See* DOJ and DHS et al., *Joint Statement on Enforcement of Civil Rights, Fair Competition, Consumer Protection, and Equal Opportunity Laws in Automated Systems* (Apr. 4, 2024).

20 *See* OECD, *Guidelines on the Protection of Privacy and Trans-border Flows of Personal Data* (Sept. 9, 2013), https://doi.org/10.1787/9789264196391-en.

21 *See* NTIA, *Dual-Use Foundation Models with Widely Available Model Weights Report* (July 30, 2024), https://www.ntia.gov/issues/artificial-intelligence/open-model-weights-report; CISA, With Open Source Artificial Intelligence, Don't Forget the Lessons of Open Source Software (July 29, 2024), https://www.cisa.gov/news-events/news/open-source-artificial-intelligence-dont-forget-lessons-open-source-software.

22 *See* CISA, *Software Must Be Secure by Design and Artificial Intelligence Is No Exception* (Aug. 18, 2023), https://www.cisa.gov/news-events/news/software-must-be-secure-design-and-artificial-intelligence-no-exception; NIST, *Guidelines for Evaluating and Red-Teaming Generative AI Models and Systems and Dual Use Foundation Models* (forthcoming).

23 *See* Helen Toner & Timothy Fist, *Regulating the AI Frontier: Design Choices and Constraints*, Georgetown University Center for Security and Emerging Technology (Oct. 26, 2023), https://cset.georgetown.edu/article/regulating-the-ai-frontier-design-choices-and-constraints/.

24 "Red-teaming" means "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." The White House, *Executive Order 14110*, *supra* note 3, § 3(d).

25 *See* L. Koessler, *Risk Thresholds for Frontier AI*, arXiv (2024), https://arxiv.org/pdf/2406.14713.

26 *See* NIST, *Managing Misuse Risk for Dual-Use Foundation Models, Initial Public Draft*, *supra* note 14.

27 *See* DHS, *Report on Reducing the Risks at the Intersection of AI and Chemical, Biological, Radiological, and Nuclear Threats* (June 2024), https://www.dhs.gov/sites/default/files/2024-04/24_0429_cwmd-dhs-fact-sheet-ai-cbrn.pdf.

28 Federal policy considers AI developers and compute infrastructure providers as themselves critical infrastructure entities within the IT sector; this Framework distinguishes between AI developers and compute infrastructure providers separately from critical infrastructure entities in order to address these entities' responsibilities on the issue of AI safety and security, and it covers the responsibilities of AI developers and critical infrastructure in Section IV.C.II and Section IV.C.III, respectively.

29 *See* ISO & IEC, *ISO 27001 Information Security, Cybersecurity, and Privacy Protection* (2022), https://www.iso.org/obp/ui/#iso:std:iso-iec:27001.

30 *See* CISA & NSA, *Joint Guidance on Deploying AI Systems Securely* (Apr. 2024), https://media.defense.gov/2024/Apr/15/2003439257/-1/-1/0/CSI-DEPLOYING-AI-SYSTEMS-SECURELY.PDF.

31 *See* CISA, *Secure by Demand Guide: How Software Customers Can Drive a Secure Technology Ecosystem* (Aug. 2024), https://www.cisa.gov/resources-tools/resources/secure-demand-guide.

32 For more on ensuring a robust deployment environment, see CISA & NSA, *supra* note 30.

33 *See* CISA, *Secure by Design: Shifting the Balance of Cybersecurity Risk: Principles and Approaches for Secure by Design Software* (Oct. 25, 2023), https://www.cisa.gov/resources-tools/resources/secure-by-design; CISA, *Cross-Sector Cybersecurity Performance Goals* (Mar. 2023), https://www.cisa.gov/cross-sector-cybersecurity-performance-goals; CISA & NSA, *supra* note 30.

34 *See* NIST, *Artificial Intelligence Risk Management Framework Playbook, Measure* 1.2 (2022), https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook/Measure.

35 *See* OECD, *Advancing Accountability in AI* (Feb. 2023), https://www.oecd.org/en/publications/2023/02/advancing-accountability-in-ai_753bf8c8.html; NIST, *The NIST Definition of Cloud Computing, SP* 800-145 (Sept. 2011), https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf.

36 The Office of Management and Budget (OMB) has provided guidance to Federal agencies on appropriately managing risks related to acquiring AI. *See* OMB, M-24-10, *supra* note 2.