TLP:CLEAR

Principles for the Secure Integration of Artificial Intelligence in Operational Technology

Publication: December 3, 2025

U.S. Cybersecurity and Infrastructure Security Agency
Australian Signals Directorate's Australian Cyber Security Centre
U.S. National Security Agency's Artificial Intelligence Security Center
U.S. Federal Bureau of Investigation
Canadian Centre for Cyber Security
German Federal Office for Information Security
Netherlands National Cyber Security Centre
New Zealand National Cyber Security Centre
United Kingdom National Cyber Security Centre

TLP:CLEAR

Table of Contents

Introduction

Since the public release of ChatGPT in November 2022, artificial intelligence (AI) has been integrated into many facets of human society. For critical infrastructure owners and operators, AI can potentially be used to increase efficiency and productivity, enhance decision-making, save costs, and improve customer experience. Despite the many benefits, integrating AI into operational technology (OT) environments that manage essential public services also introduces significant risks—such as OT process models drifting over time or safety-process bypasses—that owners and operators must carefully manage to ensure the availability and reliability of critical infrastructure.

This guidance—co-authored by the Cybersecurity and Infrastructure Security Agency (CISA) and Australian Signals Directorate's Australian Cyber Security Centre (ASD's ACSC) in collaboration with the National Security Agency's Artificial Intelligence Security Center (NSA AISC), the Federal Bureau of Investigation (FBI), the Canadian Centre for Cyber Security (Cyber Centre), the German Federal Office for Information Security (BSI), the Netherlands National Cyber Security Centre (NCSC-NL), the New Zealand National Cyber Security Centre (NCSC-NZ), and the United Kingdom National Cyber Security Centre (NCSC-UK), hereafter referred to as the "authoring agencies"—provides critical infrastructure owners and operators with practical information for integrating AI into OT environments. This guidance outlines four key principles critical infrastructure owners and operators can follow to leverage the benefits of AI in OT systems while reducing risk:

1. **Understand AI.** Understand the unique risks and potential impacts of AI integration into OT environments, the importance of educating personnel on these risks, and the secure AI development lifecycle.
2. **Consider AI Use in the OT Domain.** Assess the specific business case for AI use in OT environments and manage OT data security risks, the role of vendors, and the immediate and long-term challenges of AI integration.
3. **Establish AI Governance and Assurance Frameworks.** Implement robust governance mechanisms, integrate AI into existing security frameworks, continuously test and evaluate AI models, and consider regulatory compliance.
4. **Embed Safety and Security Practices Into AI and AI-Enabled OT Systems.** Implement oversight mechanisms to ensure the safe operation and cybersecurity of AI-enabled OT systems, maintain transparency, and integrate AI into incident response plans.

The authoring agencies encourage critical infrastructure owners and operators to review this guidance and action the principles so they can safely and securely integrate AI into OT systems.

Important Terminology

The scope of this guidance specifically covers how critical infrastructure owners and operators can help ensure the safety and security of AI systems in OT environments. As such, the authoring agencies use the following specific definitions for terms in this guidance in order to avoid conflation with their definitions in other contexts:

- **Artificial intelligence (AI)** is a system that uses machine- and human-based inputs to make predictions, recommendations, or decisions influencing real or virtual environments.[1]
- **Safety** refers to physical safety (formally, functional safety) in an OT environment. OT systems control physical systems that can harm people or property, such as systems that deliver biological or chemical agents, control operations for a dam or wastewater treatment, or automate the flow of vehicle traffic. In this guidance, "safety" as a word on its own always refers to functional safety.
- **Security** (used interchangeably in this guidance with "information security" and "cybersecurity") refers to ensuring the security properties—such as confidentiality, integrity, and availability—of information and information systems.

See **Appendix: Terminology** for a full list of definitions within the scope of this guidance and sources for these definitions.

Scope

Machine learning (ML), statistical modeling, and algorithmic calculations are all subsets of AI techniques that have been used in critical infrastructure engineering processes for many years. While ML and traditional statistical modeling are both used for predicting outcomes or making decisions based on data, they differ in their approach, assumptions, applications, and considerations for secure integration with OT systems. The scope of this guidance focuses on ML- and large language model (LLM)-based AI and AI agents because integrating OT with these types of AI systems involves more complex safety and security considerations. However, this guidance may also be applied to systems augmented with traditional statistical modeling and other logic-based automation. The following subsections define these different AI techniques.

Types of AI Techniques

Traditional statistical modeling uses mathematical formulas to accurately describe the relationships between variables. It assumes that the data follows certain distributions and that the relationships are either linear or can be approximated by linear models. Statistical modeling uses techniques such as regression analysis, hypothesis testing, and confidence intervals to directly estimate model parameters and make predictions. It is commonly used for tasks such as forecasting, optimization, and assisting in operator decision-making. Non-machine-learning-based AI systems employ algorithms to automate decision-making and control processes; in OT systems, this includes ladder logic automation routines and a class of safety instrumented systems.

Machine learning systems use algorithms to learn from data and make predictions or decisions without being explicitly programmed. The ML model can handle complex relationships and non-linear interactions between variables. ML models use various techniques—such as supervised, unsupervised, and reinforcement learning—when developing representations and making predictions based on data. ML is

[1] This document uses this AI definition from 15 U.S.C. 9401(3); however, definitions of AI may vary among groups and jurisdictions.

commonly used in fields like computer vision, natural language processing, and robotics for tasks such as image classification, speech recognition, and autonomous driving.

Large language models are advanced ML models designed to understand a natural language prompt and generate a response that humans can understand. LLMs use patterns in language and multimodal datasets in the production of complex responses to user prompts. LLM engineers usually build in randomness when generating outputs[2] so that the LLMs don't always produce the same response to the same inputs. LLMs can power generative AI applications that support critical infrastructure entities by enhancing decision-making, automating routine tasks, and optimizing maintenance schedules, with the goal of improving efficiency and reliability in operations.

AI agents are a type of software that can process data, perform decision-making capabilities, and initiate autonomous actions using AI and ML models. There are many types of agentic AI systems, including systems that use LLMs to power generative AI applications or agents and systems that combine different ML techniques, perspectives of analysis, decision-making methodologies, and autonomous action capabilities. Like LLMs, they can enhance decision-making, automate routine tasks, and optimize maintenance schedules, which enables them to improve and streamline critical infrastructure operations. Implementing error-checking can improve AI agent's performance by avoiding problems and ensuring its outputs are within the expected bounds.

AI Applications According to the Purdue Model

The Purdue Model is still a widely accepted framework for understanding the hierarchical relationships between OT and IT devices and networks. **Table 1** shows examples of established and potential AI applications in critical infrastructure according to the Purdue Model.[3] ML techniques, such as predictive models, are typically used in operational layers (0–3), while LLMs are typically used in the business context (4–5), potentially on data exported from the OT network.

Table 1. AI Applications According to the Purdue Model

Level	Description	Example AI Uses
Level 0: Field Devices	Sensors, actuators, and other devices that interact with physical processes.	**OT data source:** Field devices may generate OT data that can be used for training AI models (primarily predictive ML models) or identifying significant deviations.

[2] Sander Shulhoff, "Basic LLM Settings," *Learn Prompting*, last modified March 10, 2025, https://learnprompting.org/docs/intermediate/configuration_hyperparameters.

[3] The version of the Purdue Model used in this guidance was sourced from Manuel Humberto Santander Pelaez, "Controlling Network Access to ICS Systems," *Diaries* (blog)*, SANS Technology Institute Internet Storm Center*, July 3, 2023, https://isc.sans.edu/diary/30000.

Level	Description	Example AI Uses
Level 1: Local Controllers	Apparatus and systems designed to offer automated regulation of a process, cell, or line; examples include programmable logic controllers (PLCs) and remote terminal units (RTUs).	**AI for local control:** Some modern PLCs or edge controllers execute lightweight, pre-trained predictive models for classification for tasks like local anomaly detection, load balancing, and maintaining a known safe state.
Level 2: Local Supervisory	Observation and managerial oversight for an individual process, line, or cell; examples include supervisory control and data acquisition (SCADA) systems, distributed control systems (DCSs), and human-machine interfaces (HMIs).	**Quality control:** AI models (primarily predictive ML models) may be used for analyzing data from the SCADA system or DCS to detect early signs of equipment anomalies and alert operators that corrective action may be required.
Level 3: Site-Wide Supervisory	Monitoring, supervisory, and operational support for all or part of the regions covered by the company; examples include manufacturing execution systems and historians.	**Predictive maintenance:** AI models (primarily predictive ML models) may be used for analyzing aggregated historian OT data and predicting equipment maintenance requirements. **Support operator decision-making:** AI models may also be integrated into local supervisory systems to provide system recommendations that support operator decision-making, such as operations measurement.
Levels 4 & 5: Enterprise & Business Networks	IT systems that manage business and corporate processes and decisions; in the context of critical infrastructure and OT, examples include OT data analysis and autonomous defense for both OT and IT systems.	**Workflow optimization:** AI systems (including AI agents and LLMs) may be used for improving business processes, such as the intersection between business use cases and engineering. **Behavioral analytics and profiling of OT and IT data:** AI can be used for analyzing OT data in conjunction with IT data to measure operations, perform anomaly and threat detection, determine hardening mitigations, and provide information that supports prioritized resiliency decisions.

Principles for the Secure Integration of AI in OT

Principle 1 – Understand AI

1.1 Understand the Unique Risks of AI and Potential Impact to OT

The following section discusses AI integration risks and the potential impact to OT operations. **Table 2** provides a broad overview of known AI risks that critical infrastructure owners and operators should consider. (**Note:** This is a non-exhaustive list; critical infrastructure owners and operators should investigate risks specific to their organization.) Subsequent sections of this guidance discuss mitigation considerations for these risks; see cross-references in the Mitigations column of **Table 2**.

Table 2. AI Risks and Impacts in an OT Environment

AI Risks in an OT Environment	OT Impacts	Mitigations
Cybersecurity Risks: AI data, models, and deployment software can be manipulated to cause incorrect outcomes or bypass security and functional safety measures or guardrails. Traditional cybersecurity risks remain within AI systems; as such, security measures like access control, auditing, and encryption still apply for securing AI and AI-enabled systems. In addition, AI-enabled systems are subject to AI-specific cybersecurity risks, such as prompt injection.	Impacted system availability, functional safety risks, financial losses, reputational damage, network/OT compromise, cascading compromise.	1.2 Understand the Secure AI System Development Lifecycle 2.4 Evaluate Challenges in AI-OT System Integration 3.3 Conduct Thorough AI Testing and Evaluation
Data Quality: AI models can only be as effective as the quality of their training data. Collecting high-quality, normalized sensor data can be difficult, especially in distributed OT environments. Centralizing this operational data creates its own risk as threat actors can use it to create a more targeted engineering impact.	Reduced OT safety and system availability from poor data quality.	2.2 Manage OT Data Security Risks for AI Systems

AI Risks in an OT Environment	OT Impacts	Mitigations
AI Model Drift: AI models may become less accurate over time due to data being introduced to the model that is not represented by the model's initial training data. Alterations to production processes can affect model performance.	Increased dependencies on changes, loss of productivity, reduced OT safety and system availability.	4.1 Establish Monitoring and Oversight Mechanisms for AI in OT
Lack of Explainability: Understanding an AI model's decision-making process may be difficult; this makes it challenging to diagnose and correct errors or properly audit a system.	Increased recovery time, functional safety risks, reduced system availability, complexity in troubleshooting.	1.3 Educate Personnel on AI 4.1 Establish Monitoring and Oversight Mechanisms for AI in OT
Operator Cognitive Load and Unnecessary Downtime: AI may generate alarm errors that could cause unnecessary downtime or safety incidents. These alarm errors increase cognitive load, distract operators, and potentially lead to further human error.	Reduced system availability, functional safety risks, financial losses, reputational damage.	1.3 Educate Personnel on AI 4.1 Establish Monitoring and Oversight Mechanisms for AI in OT
Regulatory Compliance: Compliance with regulatory requirements, such as those related to OT safety or privacy, can be challenging due to the evolving nature of AI, technical standards, and regulatory frameworks. For example, while producing a robust audit trail of AI-driven decision-making may be difficult, it may be required for regulatory compliance.	Functional safety risks, financial losses, reputational damage.	3.4 Navigating Regulatory and Compliance Considerations for AI in OT 4.1 Establish Monitoring and Oversight Mechanisms for AI in OT 4.2 Embed Safety and Failsafe Mechanisms
AI Dependency: Overreliance on AI can lead to operators missing critical safety-related information if the AI misses it, and losing valuable skills for safely operating equipment manually or without the AI functionality.	Dependence on technology, complexity in troubleshooting.	1.3 Educate Personnel on AI

AI Risks in an OT Environment	OT Impacts	Mitigations
Interoperability Issues: Integrating AI systems with existing OT infrastructure can be complicated by interoperability challenges, which may arise from differences in OT communication protocols or data formats.	Increased maintenance costs, recovery challenges.	2.1 Consider the OT Business Case for AI Use 2.4 Evaluate Challenges in AI-OT System Integration 3.1 Establish Governance Mechanisms for AI in OT 3.3 Conduct Thorough AI Testing and Evaluation
Complexity: Incorporating AI usually requires increasing the complexity of the overall system to support process automation.	Functional safety risks, complexity in troubleshooting.	2.1 Consider the OT Business Case for AI Use 2.4 Evaluate Challenges in AI-OT System Integration
Reliability: AI may not be reliable enough to independently make critical decisions in industrial environments. AI can also hallucinate (i.e., fabricate a plausible, but false, response or data), which would provide operators with incorrect information for decision-making. As such, AI such as LLMs almost certainly should not be used to make safety decisions for OT environments.	Decisions made by AI developers may pose OT safety and reliability risks, increased documentation costs, uncertainty due to changes in automated decision-making over time, increased risk of cascading failure due to tighter coupling of actions. False information provided to decision makers poses risks of unsafe operating conditions, equipment damage, production halts.	2.1 Consider the OT Business Case for AI Use 3.1 Establish Governance Mechanisms for AI in OT 3.2 Integrating AI Into Existing Security and Cybersecurity Frameworks 3.3 Conduct Thorough AI Testing and Evaluation 4.1 Establish Monitoring and Oversight Mechanisms for AI in OT 4.2 Embed Safety and Failsafe Mechanisms

1.2 Understand the Secure AI System Development Lifecycle

To address the unique challenges of integrating AI into OT environments, critical infrastructure owners and operators should verify that the AI system was designed securely and understand their roles and responsibilities through the AI system's lifecycle. Similar to hybrid ownership models used with cloud systems, owners and operators must clearly define and communicate these roles and responsibilities with the AI system manufacturer, OT supplier, and any system integrator or managed service provider roles.

NCSC-UK and CISA's joint Guidelines for Secure AI System Development emphasizes the following key stages of the AI system development lifecycle:[4]

- **Secure Design.** Design the AI system with security considerations in mind from its inception, including using robust coding, protocols, and data protection measures.
- **Secure Procurement or Development.** Select vendors who adhere to secure practices and develop AI systems using secure methodologies and tools.
- **Secure Deployment.** Deploy the AI system using methods that maintain its security posture, including using proper network segmentation and access control, as well as verifying and validating that the AI system works as intended.
- **Secure Operation and Maintenance.** Ensure the AI system continues operating securely throughout its lifecycle, including by implementing regular updates and patches, and monitoring potential vulnerabilities.

Critical infrastructure owners and operators should also carefully evaluate the trade-offs between different methods for sourcing an AI system:

- **Procure an AI System.** Select a pre-developed AI system from a vendor that meets specific security requirements and that the OT supplier agrees with.
- **Develop an AI System.** Build an AI system in house; this enables complete control over its design and implementation.
- **Customize an Existing AI System.** Work with a vendor to tailor their existing AI system to meet specific OT system needs.

Where possible, critical infrastructure owners and operators should demand AI systems that are secure by design and will not negatively impact OT operation and safety. Critical infrastructure owners and operators should consult CISA's Secure by Design webpage and resources, and the joint guidance Secure by Demand: Priority Considerations for Operational Technology Owners and Operators when Selecting Digital Products for opportunities to incorporate these principles into the design of their AI and OT systems.

1.3 Educate Personnel on AI

Integrating AI into OT environments can lead to personnel relying too much on automation, resulting in reduced human oversight and situational awareness. This can have significant consequences, including:

- **Dependency Risks and Skill Erosion.** Heavy reliance on AI may cause OT personnel to lose manual skills needed for managing systems during AI failures or system outages.
- **Skill Gaps.** OT personnel may misinterpret AI outputs, leading to incorrect actions; OT personnel may also lack expertise for managing or troubleshooting AI systems if they malfunction.

[4] The UK Government's Code of Practice for the Cyber Security of AI and its technical implementation guide also provide scenario-based cybersecurity mitigation advice according to the secure AI system development lifecycle.

Critical infrastructure owners and operators may mitigate these risks by focusing on skill development and cross-disciplinary collaboration, such as:

- **Training OT teams** on AI fundamentals and threat modeling so teams can effectively interpret and validate AI outputs and maintain operational competencies alongside AI systems—for example, training teams to use alternative sensors (e.g., human senses, vibration or temperature sensors, voltage readings) for validating AI output—and know what actions to take if AI outputs are invalid.
- **Developing clear standard operating procedures (SOPs)** for all operations (including AI-related operations), interventions, and incidents to support stakeholder awareness of their roles and responsibilities in managing AI-enabled OT systems.
- **Leveraging explainable AI** by having operators request that AI outputs include clear and transparent documentation of decision-making processes; this enables humans to better understand and validate outputs.

Principle 2 – Consider AI Use in the OT Domain

2.1 Consider the OT Business Case for AI Use

Before incorporating an AI system into their OT environment, critical infrastructure owners and operators should assess if AI technologies are the most appropriate solution for their specific needs and requirements compared to other technologies. Critical infrastructure owners and operators should further consider whether an established capability meets their needs before pursuing more complex and novel AI-enabled solutions. While AI comes with unique benefits, it is an evolving technology that requires continuous evaluation of risks.

This assessment should incorporate various factors—including security, performance, complexity, cost, and effect on OT-environment safety depending on the specific application—and assess the benefits and risks of using the AI technologies against the functional requirements the application should meet. Critical infrastructure owners and operators should understand the organization's current capacity for maintaining an AI system in their OT environment and the potential impact of expanding the environment's risk surface, such as requiring additional hardware and software for processing data through models or additional security infrastructure to protect the expanded attack surface.

If the assessment indicates an AI system is the best solution, then critical infrastructure owners and operators should follow the secure AI system development lifecycle outlined above and consult AI risk management frameworks, such as the National Institute of Standards and Technology's (NIST's) AI Risk Management Framework to help ensure the system is used safely and securely.

Example Assessment of an AI in OT Business Use Case

The following example shows a fictional assessment by a critical infrastructure organization on the feasibility of using AI for supporting predictive maintenance for an industrial generator. The assessment includes performance thresholds and organizational safety and security requirements that technology must meet in order for critical infrastructure owners and operators to recommend using AI in the OT environment.

Use Case: Use an AI system for predictive maintenance for an industrial generator.

Problem Statement: High downtime and maintenance costs associated with industrial generator failures.

Goal: Implement an AI-powered predictive maintenance solution that detects potential generator failures 30 days in advance.

Risk: If AI does not perform predictive maintenance correctly, equipment could be replaced prematurely.

Key Stakeholders:

- Operations team responsible for scheduling generator maintenance.
- Maintenance personnel who perform generator repairs and replacements.
- Equipment owners who are impacted by generator downtime and maintenance costs.

Requirements:

- Historical data access on generator performance, including sensor readings and maintenance records.
- Ability to process large OT datasets in near-real time.
- Reliable predictive failure detection.

Corresponding Safety and Security Requirements:

- Aggregated data is protected in transit and at rest, logging any access or changes.
- Data aggregation does not exceed 80% network bandwidth threshold.
- Erroneous alarms are rare and provide context, remaining actionable and preventing operator fatigue.

Success Metrics:

- Reduction in downtime by 25%.
- Decrease in maintenance costs by 15%.
- Improvement in overall equipment effectiveness by 10%.

2.2 Manage OT Data Security Risks for AI Systems

Data-Related Challenges

When integrating AI into OT environments, critical infrastructure owners and operators should work with AI model developers when addressing several data-related challenges, including:

Data Assurance. Understand where OT data used for training AI models is stored and ensure it is within the organization's control. Securely manage access to OT data, including who can view, access, or modify it. Understand how AI vendors access and use the organization's OT data, especially if it involves remote, cloud, or offshore access.

Data Sovereignty. Be aware that businesses based in foreign countries are subject to foreign government control and laws and may have to comply with directions that may be contrary to your business' interests.

Exposure of Sensitive Information. Minimize risk by not sharing sensitive data with AI models, especially if the AI models are hosted in an environment hosted or controlled by external parties, such as public cloud infrastructure.

Data Privacy and Security. Protect proprietary and personal information within OT datasets, including by instituting protection from access abuse, intentional or inadvertent data poisoning, or dependency on synthetic, generated data.

Data Silos. Address the complexity of integrating AI systems in OT environments due to OT/IT network segregation, proprietary protocols, formats, diversity of OT products, and a multitude of suppliers.

Data Quality and Availability. Apply specialized domain knowledge that curates high-quality, comprehensive datasets for effective AI performance. This can be challenging in industrial settings; potential difficulties include the environment containing proprietary or outdated systems and bespoke solutions, and the difficulty of capturing safety-related edge cases. OT operator expertise is necessary for capturing this specialized knowledge. Additionally, operators should work with AI model developers to ensure a data integrity program covers AI systems.

Prioritizing OT Data Protection

Prioritize the protection of critical types of OT data, including the following:

Engineering Configuration Data. These include network diagrams, asset inventories, documentation on operations sequences, safety-related information, logic diagrams, and schematics. These data points have enduring value and are highly valuable to cyber adversaries.

Ephemeral OT Data. Data from industrial measurement technology, especially process measurement technology (e.g., voltage or temperature, pressure levels, mass/volume flow rates) can provide insight into organizational activities or system behavior. If that data is used to train or update an AI model, the data may become accessible or stored (statistically) for a longer period of time in the model. As such, securing these data points can be important for protecting intellectual property (IP) and patterns of activity.[5]

2.3 Understanding the Role of OT Vendors in AI Integration

OT vendors play a crucial role in advancing AI integration into OT environments. Some OT devices now come with built-in AI technology, which may require internet connectivity to function.

Emerging trends among OT vendors include:

- **Operator-Facing AI.** OT vendors increasingly integrate AI capabilities directly into their devices, such as models used for predicting grid frequency dynamics.
- **Intelligent Devices.** As technology advances, expect the emergence of increasingly sophisticated "intelligent devices" that have included AI capabilities for engineering and modifying control.

[5] For further information on the protection of OT data, see ASD's ACSC's Principles of Operational Technology Cybersecurity and NSA's joint guidance AI Data Security: Best Practices for Securing Data Used to Train & Operate AI Systems.

Critical infrastructure owners and operators should demand transparency and security considerations from OT vendors regarding how AI technologies are embedded into their products. This includes:

- **Contractual Agreements.** Negotiate contractual agreements that ensure OT vendors provide explicit details on AI features and functionalities.
- **Software Supply Chain and Software Bill of Materials (SBOM).** Require OT vendors to provide information on how AI is embedded into their OT products (including by requesting vendors provide an SBOM)[6] and clarify the supply chain of AI models used in their products (e.g., where models are hosted).
- **Vendor Notifications.** In addition to typical vulnerability disclosure policies for the lifecycle of the product, the AI vendor should disclose if they have discovered an indication that the AI can provide improper advice or take inappropriate action.
- **Explicit Data Usage Policy Review and Enforcement.** Operators may not want vendors to train AI on operational data as it could include IP or be sensitive. Operators can control their information via a data usage policy with clear reference to data residency, communications paths, encryption, and storage.
- **Increased Connectivity.** Ask if vendors can operate on premises or without continuous access to the internet/vendor cloud.
- **Disabling AI Features.** Specify conditions under which certain AI features can be disabled or enabled, with control given to the operator.

By demanding transparency and control from OT vendors, critical infrastructure owners and operators can better manage the risks associated with embedded AI systems in OT products.

2.4 Evaluate Challenges in AI-OT System Integration

When integrating AI into OT environments, critical infrastructure owners and operators should carefully evaluate the existing infrastructure to ensure compatibility and security.

Challenges in AI-OT System Integration

Some challenges an organization may face when integrating AI into OT systems include:

- **Increased System Complexity and Vulnerabilities.** AI integration adds complexity to OT systems, potentially creating new attack surfaces and vulnerabilities that malicious actors can exploit—this is a particularly important consideration when AI introduces new, remotely accessible, internet-exposed attack paths.
- **Cloud Security Risks.** Integrating AI into cloud supervisory control and data acquisition (SCADA) environments may introduce additional cybersecurity risks or cause data transmission latency.

[6] AI systems are software systems, and the minimum elements of an SBOM for AI software are the same as the minimum elements for all software; see CISA's 2025 Minimum Elements for a Software Bill of Materials. A Shared G7 Vision on Software Bill of Materials for Artificial Intelligence, produced by the G7 Cybersecurity Working Group, discusses potential additional elements for describing AI systems.

- **Forward and Reverse Compatibility.** OT system design evolution may be necessary to accommodate the secure integration of AI, as many OT environments rely on older equipment that lacks standardized data formats and computing, complicating AI data integration and analysis.
- **Latency and Real-Time Constraints.** AI systems may not meet the strict timing requirements of OT environments; these requirements vary significantly by sector and matter more if the AI system is actively involved in the control process.
- **AI Vendor Transparency.** Lack of vendor transparency could prevent insight into functions that make external connections or modify standard engineering workflows.

To mitigate these challenges, critical infrastructure owners and operators should:

- **Integrate AI systems into their overall security and cybersecurity framework** (see **3.2 Integrating AI Into Existing Security and Cybersecurity Frameworks**).
- **Add AI security considerations to a comprehensive security strategy** that also includes traditional cybersecurity considerations, such as data encryption, access controls, and intrusion detection systems.
  - Critical infrastructure owners and operators should define and validate security clauses in cloud contracts, explicitly outlining any AI security responsibilities, compliance standards, and support provisions, including data protection, access controls, incident response, and audit capabilities.
  - Cloud providers should provide detailed documentation that outlines security obligations specific to AI capabilities, in addition to traditional cloud security shared responsibility models.
- Consider existing OT infrastructure and **assess and develop an integration plan for AI systems**.
  - Consider using test infrastructure before deployment to production systems, if possible (see **Principle 3 – Establish AI Governance and Assurance Frameworks**).
- **Encourage push-based architectures** where data is pushed out of the OT network for AI systems to use without persistent access into the OT network.[7]
- **Prioritize the organization's control over critical functions that AI systems may integrate with or enable** when hosting AI systems locally or in the cloud.
  - Ensure there are failsafe mechanisms that revert to traditional automation or manual for any AI-enabled system processes.
- Integrate AI systems the same as any new OT systems: **test AI systems for safety impacts (e.g., latency, interoperability) and verify they work within existing device management policy** (i.e., new connection paths such as remote access are approved and work through the existing demilitarized zone [DMZ] or jump host infrastructure).

[7] Visit ASD's ACSC's Principles of Operational Technology Cybersecurity for further information.

- **Limit active control of OT infrastructure by AI without a human in the loop** to account for safety concerns and latency limitations.
- **Regularly update and validate AI models for accuracy and effectiveness.**

Principle 3 – Establish AI Governance and Assurance Frameworks

3.1 Establish Governance Mechanisms for AI in OT

Effective governance structures are essential for the safe and secure integration of AI into OT environments. This involves establishing clear policies, procedures, and accountability structures for AI decision-making processes within OT. An AI governance structure should include the key stakeholders listed below, as well as any AI vendors needed for maintaining oversight during procurement, development, design, deployment, and operations.

Key Stakeholders in AI Governance Mechanisms

Leadership. Securing commitment from senior leadership, including the CEO and CISO, is essential for establishing a robust AI governance framework. This helps ensure that the organization's leadership is fully invested in the secure lifecycle management of AI systems and considers AI security risks and mitigations alongside AI functionality.

OT/IT Subject Matter Experts. Engaging OT, IT, and AI subject matter experts is critical for effective and secure integration of AI systems into OT environments. These experts provide valuable insights into the OT environment and can help identify potential risks and challenges associated with AI integration.

Cybersecurity Teams. Collaborating with cybersecurity teams is vital for developing policies and procedures that protect sensitive OT data used by AI models. Cybersecurity teams can help identify potential vulnerabilities and provide mitigation recommendations to help maintain the security of the organization's data.

Additional Components in AI Governance Mechanisms

Other key components of governance structures may include:

- **Enforcing strict data governance policies** that protect sensitive OT data used by AI models, including encryption, access controls, and user behavior analytics.
- **Establishing clear roles and responsibilities** that ensure everyone involved in the development, deployment, and operations and maintenance of AI systems (e.g., data owners, model developers, and end users) understands their tasks and expectations—and to avoid liability and confusion over stakeholder responsibilities in the event of safety or operational incidents.
- **Implementing regular audits and compliance testing** to help identify potential issues and ensure ongoing adherence to AI governance requirements.
- **Continuously validate and verify the performance of AI systems** to make sure they meet the organization's objectives and regulatory requirements.

3.2 Integrating AI Into Existing Security and Cybersecurity Frameworks

When integrating AI into OT environments, critical infrastructure owners and operators should consider the existing security and cybersecurity frameworks that govern these systems and embed AI system assessments within existing risk evaluation, mitigation, and monitoring processes. This means that traditional cybersecurity requirements, vulnerability management, and critical infrastructure regulations must be factored in when integrating AI systems. These processes include:

- **Regular Security Audits and Risk Assessments.** Conduct, or obtain proof of the AI vendor conducting, regular security audits and risk assessments to identify potential vulnerabilities in AI systems.
- **Robust Security Controls.** Implement robust security controls, such as encryption, access controls, and intrusion detection systems, that protect and detect anomalies in AI systems and data.[8]
  - Collect flow logs and access logs for AI endpoints and track data egress by asset and identity.
  - Integrate with data loss prevention for prompts and output inspection.
- **AI-Tailored Security Information.** Security teams should incorporate AI-related tactics, techniques, and procedures (TTPs) when evaluating risk or modeling threats. For instance, when using the MITRE ATT&CK® Matrix for Enterprise for threat actor behavior mapping, teams should also incorporate AI-related TTPs using tools such as the MITRE Adversarial Threat Landscape for Artificial-Intelligence Systems (ATLAS™) Matrix, which is tailored for TTPs against AI-enabled systems,

3.3 Conduct Thorough AI Testing and Evaluation

Thorough testing and evaluation (T&E) are crucial when introducing AI into OT environments to support the safe and reliable operation of these systems. Operators should initially conduct tests of the AI system on infrastructure specifically designed for testing. Low fidelity testing can allow for faster iterations of testing early in the T&E process. Alternatively, work with vendors to understand their testing and if it includes dependencies (e.g., operating system versions, protocols). As the system matures, operators can test with more realistic, non-production systems, including hardware in the loop.

Operators should only move the AI system into production for additional testing after sufficient testing in a non-production environment. Virtualized controllers can speed up this testing process when physical effects do not need modelling.

Critical infrastructure owners and operators should also comply with traditional data protection mechanisms when conducting AI testing and evaluation, such as avoiding production data exposure in non-production environments.

[8] See NSA's joint Cybersecurity Information Sheet: AI Data Security and NCSC-UK's joint Guidelines for Secure AI System Development for more information on using controls for protection and anomaly detection in AI systems and data.

3.4 Navigating Regulatory and Compliance Considerations for AI in OT

As more critical infrastructure owners and operators integrate AI technologies into their OT environments, regulatory and compliance considerations are a key challenge. Some examples include:

- **Lack of OT-Oriented AI Standards.** Current international AI technical standards are broadly aligned to the deployment of AI systems into IT environments.
- **Auditability.** Tracing or explaining AI decisions can be difficult, which can complicate regulatory audits.
- **Safety Certifications.** AI systems may not meet the rigorous safety standards required in critical infrastructure and OT environments.

Critical infrastructure owners and operators should evaluate the applicability of current AI technical standards in their OT domain as AI technical standards are rapidly evolving. Top AI technical standards from the European Telecommunications Standards Institute's (ETSI's) Technical Committee Securing Artificial Intelligence are outlined below:

- ETSI TR 104 128 Securing Artificial Intelligence (SAI); Guide to Cyber Security for AI Models and Systems
- ETSI TS 104 223 Securing Artificial Intelligence (SAI); Baseline Cyber Security Requirements for AI Models and Systems
- ETSI TR 104 048 Securing Artificial Intelligence (SAI); Data Supply Chain Security

Critical infrastructure owners and operators should continuously validate and verify that the performance of AI systems meets stringent OT performance and safety regulations. Critical infrastructure owners and operators should also identify and deploy thresholds for defaulting back to non-AI systems in OT, such as if AI system outputs fall below performance and safety thresholds.

Principle 4 – Embed Oversight and Failsafe Practices Into AI and AI-Enabled OT Systems

Ultimately, humans are responsible for functional safety. Humans make tools that ensure or operationalize oversight, safety, and failsafe practices—this is no different for AI tools.

4.1 Establish Monitoring and Oversight Mechanisms for AI in OT

Critical infrastructure owners and operators should implement oversight of AI-enabled OT systems by taking inventory of any AI components, as well as other components reliant on the AI. Log and monitor inputs and outputs for these components. Also, establish and maintain a known good state or thresholds for safe behavior in an OT environment, allowing for knowledge of when maintenance or restoration should be performed from a backup. Consider the following points when embedding safety processes into AI-enabled OT systems to ensure effective implementation and management:

Human-in-the-Loop Decision-Making. Provide adequate transparency that involves operators and engineers in decision-making, especially for critical OT operations and actions. For more passive AI systems, operators and engineers can implement this by incorporating the recommendations into an existing change

management process. Use caution with active AI systems directly influencing control, as problems can escalate before operators become aware of them. Where AI is actively updating control logic, use safety thresholds, alternative sensor output, or state changes that add human-in-the-loop intervention points.

Benefits of human-in-the-loop decision-making include:

- **Improved OT-Environment Safety.** Operators provide context and make informed decisions when interacting with AI-driven systems.
- **Enhanced Reliability.** Human oversight helps detect and correct potential errors or anomalies in AI performance; oversight also maintains human training, which is critical in an emergency.
- **Increased Trust.** Organizations build trust in the AI system and its decision-making processes by involving operators.

Understand the correctness of AI system results to support continued safe operation of systems in an OT environment. It is vital for critical infrastructure owners and operators to understand the states where an AI system can fail to produce accurate and reliable results. This understanding includes expectations for false positives and false negatives in the system's performance, and how the false positives compare to the base rate of true positives.

Implement anomaly detection and behavioral analytics. Establish safe operating bounds for OT devices that detect AI drift, model changes that impact safety and performance, or security risks. As operator processes mature, software safety thresholds can shift from setpoints to anomaly detection of increasingly sophisticated faults. Configure logging so AI decisions can be tracked for compliance and forensic analysis, and so the logged AI identity is distinct from any typical machine or user identifiers.

The example below demonstrates how operators and engineers should monitor a predictive maintenance system (with read-only access to OT data output) in the data zone that uses ML to produce recommendations:

- **AI System Output.** Predictive maintenance recommendations for equipment downtime.
- **Anomaly Detection Algorithm.** Statistical process control that detects outliers in predicted downtime values.
- **Monitoring Tools.** Real-time dashboards, charts, or metrics that track AI system performance and detect anomalies—ideally, these are integrated into existing human-machine interface (HMI) views for consolidated monitoring.
- **Audit Trail.** Logging of all AI system operating data (including timestamps, inputs, and outputs) for auditing and analysis of AI system behavior.
- **Offensive Security Assessments/AI Red Teaming.** Regularly evaluate AI system functions, identifying vulnerabilities and testing resilience.
- **Network and Egress Security.** Define and enforce network controls; see CISA's Cybersecurity Performance Goals (CPGs) 2.F.

Establish key performance indicators (KPIs) that measure AI effectiveness and track progress over time. Critical infrastructure owners and operators should schedule regular review sessions with AI stakeholders, such as vendors, governance boards, and operators, to discuss results, address concerns, and identify areas for improvement. Visit CISA's Artificial Intelligence webpage for more information.

Continuously validate and refine AI models in simulated environments before deployment. Regularly update threat models with AI-specific attack vectors (such as adversarial inputs or data poisoning) and monitor AI system performance for anomalies or manipulation attempts. Continuously update and refine AI models with new OT data to improve precision and reduce false positives/negatives. See NSA's joint guide Deploying AI Systems Securely for more information.

Explore new AI explainability and transparency tools. Explainable AI (XAI) and transparent AI are evolving fields of research that seek to make AI systems more understandable. **Explainability** focuses on making the reasoning behind individual AI decisions understandable to users, while **transparency** emphasizes making the overall AI system's development and operation open and accessible. Essentially, explainability clarifies *why* an AI made a specific decision, while transparency focuses on *how* the AI system works as a whole. Critical infrastructure owners and operators should, where possible, explore interpretable models or tools that make AI decisions more understandable to humans.

LLMs, predictive, or deep learning systems often operate opaquely, making auditing or understanding their decision rationale difficult. Such transparency is critical for safety and regulatory compliance in OT. XAI and transparent AI tools are designed to help AI developers understand the operation of an AI model; however, it is an open question whether these techniques would be adequate for OT environments.

Prefer push-based or brokered architectures that move required features or summaries out of OT without granting persistent inbound access. Where data must traverse to business networks, use one-way transfer patterns and audited staging buffers. This method for implementing segmentation helps operators maintain existing segmentation best practices, such that the AI system is not a persistent attack path into OT.

4.2 Embed Safety and Failsafe Mechanisms

Establish failsafe mechanisms that enable AI systems to fail gracefully without disrupting critical operations. Incorporate new AI system failure states, including how to bypass or replace an AI system, into existing functional safety and incident response processes. Integrating an AI system into existing OT networks inevitably generates new failure states for the overall critical infrastructure system. Therefore, operators responsible for revising the existing functional safety and incident response processes should incorporate these new failure states as they are critical to ensuring safe operation of these systems.

Design functional safety procedures that account for the AI system. Each critical infrastructure sector has its own safety states and procedures. Per **Principle 2 – Consider AI Use in the OT Domain**, critical infrastructure owners and operators should review how they are integrating the AI system into their existing procedures and create new safe use and implementation procedures that focus on the AI system integration into the OT environment.

Incorporate AI considerations into the cybersecurity incident response plan. Despite organizations' best efforts at mitigation, risk cannot be reduced to zero; incidents are inevitable. To account for this, critical infrastructure owners and operators should update their incident response plans and functional safety procedures with steps for responding to malicious activity directed against an AI system and AI system failure. Visit CISA's AI Cybersecurity Collaboration Playbook for more guidance on collaborating with stakeholders on AI cybersecurity risks and voluntary information sharing on AI cybersecurity incidents and vulnerabilities. As the number of deployed AI systems increases, so will the number of AI-enabled attacks on existing IT and OT systems.

Conclusion

The integration of AI into OT presents both opportunities and risks to critical infrastructure owners and operators. While AI can enhance efficiency, productivity, and decision-making, it also introduces new challenges that require careful management to support the safety, security, and reliability of OT systems. For successful mitigation of the risks of integrating AI into OT systems, it is essential critical infrastructure owners and operators follow the principles in this guidance: understand AI, consider AI use in the OT domain, establish AI governance and assurance frameworks, and embed safety and security practices into AI and AI-enabled OT systems. By adhering to these principles and continuously monitoring, validating, and refining AI models, critical infrastructure owners and operators can achieve a balanced integration of AI into the OT environments that control vital public services.

Resources

Readers may find additional information on AI, AI and OT security, and related topics discussed in this guidance in the following resources:

- CISA's Artificial Intelligence webpage
- CISA's AI Cybersecurity Collaboration Playbook
- NSA's joint guidance AI Data Security: Best Practices for Securing Data Used to Train & Operate AI Systems
- NSA's joint guidance Deploying AI Systems Securely
- NCSC-UK and CISA's joint Guidelines for Secure AI System Development
- UK Government's Code of Practice for the Cyber Security of AI and its technical Implementation Guide for the AI Cyber Security Code of Practice
- NIST's AI Risk Management Framework
- MITRE's MITRE ATLAS Matrix
- ASD's ACSC's joint guidance Principles of Operational Technology Cybersecurity
- CISA's Secure by Design webpage
- CISA's joint guidance Secure by Demand: Priority Considerations for Operational Technology Owners and Operators when Selecting Digital Products
- CISA's 2025 Minimum Elements for a Software Bill of Materials (public comment draft)
- G7 Cybersecurity Working Group's A Shared G7 Vision on Software Bill of Materials for Artificial Intelligence
- U.S. Code 15 U.S.C. 9401(3)
- European Telecommunications Standards Institute's (ETSI's) Technical Committee Securing Artificial Intelligence
- CISA's Cybersecurity Performance Goals (CPGs)

Disclaimer

CISA and the authoring agencies do not endorse any commercial entity, product, company, or service, including any entities, products, or services linked within this document. Any reference to specific commercial entities, products, processes, or services by service mark, trademark, manufacturer, or otherwise, does not constitute or imply endorsement, recommendation, or favoring by CISA and the authoring agencies.

This document does not create policies, impose requirements, mandate actions, or override existing legal or regulatory obligations. All actions taken under this document are voluntary, so anyone taking actions described in this document does so of their own volition.

Acknowledgements

BP p.l.c., Cisco, Darktrace, Fortinet, Nozomi Networks, Palo Alto Networks, and Australian critical infrastructure organizations contributed to this guidance document.

Version History

December 3, 2025: Initial version.

Appendix: Terminology

This document uses the following definitions for technical concepts. However, as definitions of AI and OT may vary among groups, readers should also understand their local jurisdiction's specific definition of these concepts and consider how they are applicable to this guidance.

Artificial Intelligence (AI). AI can be defined as:[9]

> A machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI 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 in the formulation of options for information or action.

For more definitions of AI concepts, visit ASD's ACSC's *Convoluted Layers: An Artificial Intelligence Primer*.[10]

Operational Technology (OT). NIST defines OT as:

> Programmable systems or devices that interact with the physical environment (or manage devices that interact with the physical environment). These systems/devices detect or cause a direct change through the monitoring and/or control of devices, processes, and events. Examples include industrial control systems, building management systems, fire control systems, and physical access control mechanisms.

AI-Enabled OT System. This is defined as any OT system and/or OT network that has at least one AI component within the overall system.

Risks. As defined in NIST's AI Risk Management Framework: "[R]isk refers to the composite measure of an event's probability of occurring and the magnitude or degree of the consequences of the corresponding event." In this guidance, the definition of AI risk is the measure of probability that integrating AI into a system will cause an event that harmfully impacts the safety, security, or function of critical infrastructure systems in an organization (or the ecosystem that includes the organization), and the magnitude of the consequences of this event.

While AI risk includes broader considerations across industry, academia, and other stakeholders—covering a wide range of issues, such as fairness, bias, and ethics, and the misuse of AI (including harmful content generation)—these issues are not discussed in this guidance. For a discussion of these additional types of harm, visit NIST's AI Risk Management Framework.

---

[9] This definition is taken from 15 U.S.C. § 9401(3). Executive Order 14144: Strengthening and Promoting Innovation in the Nation's Cybersecurity also uses this definition. Other jurisdictions may use their own definition for AI.

[10] See ASD's ACSC's guidance Convoluted Layers: An Artificial Intelligence Primer for additional information on AI concepts.

Safety. IEC 61508 defines functional safety in the context of Electrical/Electronic/Programmable Electronic (E/E/PE) safety-related systems, in which the safety of these systems depends on them operating correctly.[11] NIST defines safety in the context of OT security as "freedom from conditions that can cause death, injury, occupational illness, damage to or loss of equipment or property, or damage to the environment." The terms "safety" and "functional safety" are used in this document specifically in these contexts when discussing safety for AI-enabled OT systems.

Security. The terms **security**, **information security**, and **cybersecurity** have a similar meaning and are used interchangeably in this guidance when defining security for AI-enabled OT systems. NIST defines information security as protecting information and information systems from:

> [U]nauthorized access, use, disclosure, disruption, modification, or destruction in order to provide—
>
> - integrity, which means guarding against improper information [or technology] modification or destruction, and includes ensuring information non-repudiation and authenticity;
> - confidentiality, which means preserving authorized restrictions on access and disclosure, including means for protecting personal privacy and proprietary information; and
> - availability, which means ensuring timely and reliable access to and use of information [and technology].

NIST further defines security as the "establishment and maintenance of protective measures" as part of an organization's broader security policy that prevents unauthorized access and use of systems and data, ensuring continuous operability. Security also includes capabilities and processes that identify, prevent, protect against, respond to, or recover from compromises against such systems.

[11] "Overview of IEC 61508 & Functional Safety," International Electrotechnical Commission (IEC), (PowerPoint, 2022), https://assets.iec.ch/public/acos/IEC%2061508%20&%20Functional%20Safety-2022.pdf?2023040501.

References

Humberto Santander Pelaez, Manuel. "Controlling Network Access to ICS Systems." *Diaries* (blog), *SANS Technology Institute Internet Storm Center*. July 3, 2023. https://isc.sans.edu/diary/30000.

"Overview of IEC 61508 & Functional Safety." International Electrotechnical Commission (IEC), (PowerPoint, 2022). https://assets.iec.ch/public/acos/IEC%2061508%20&%20Functional%20Safety-2022.pdf?2023040501.

Shulhoff, Sander. "Basic LLM Settings." *Learn Prompting*. Last modified March 10, 2025. https://learnprompting.org/docs/intermediate/configuration_hyperparameters.