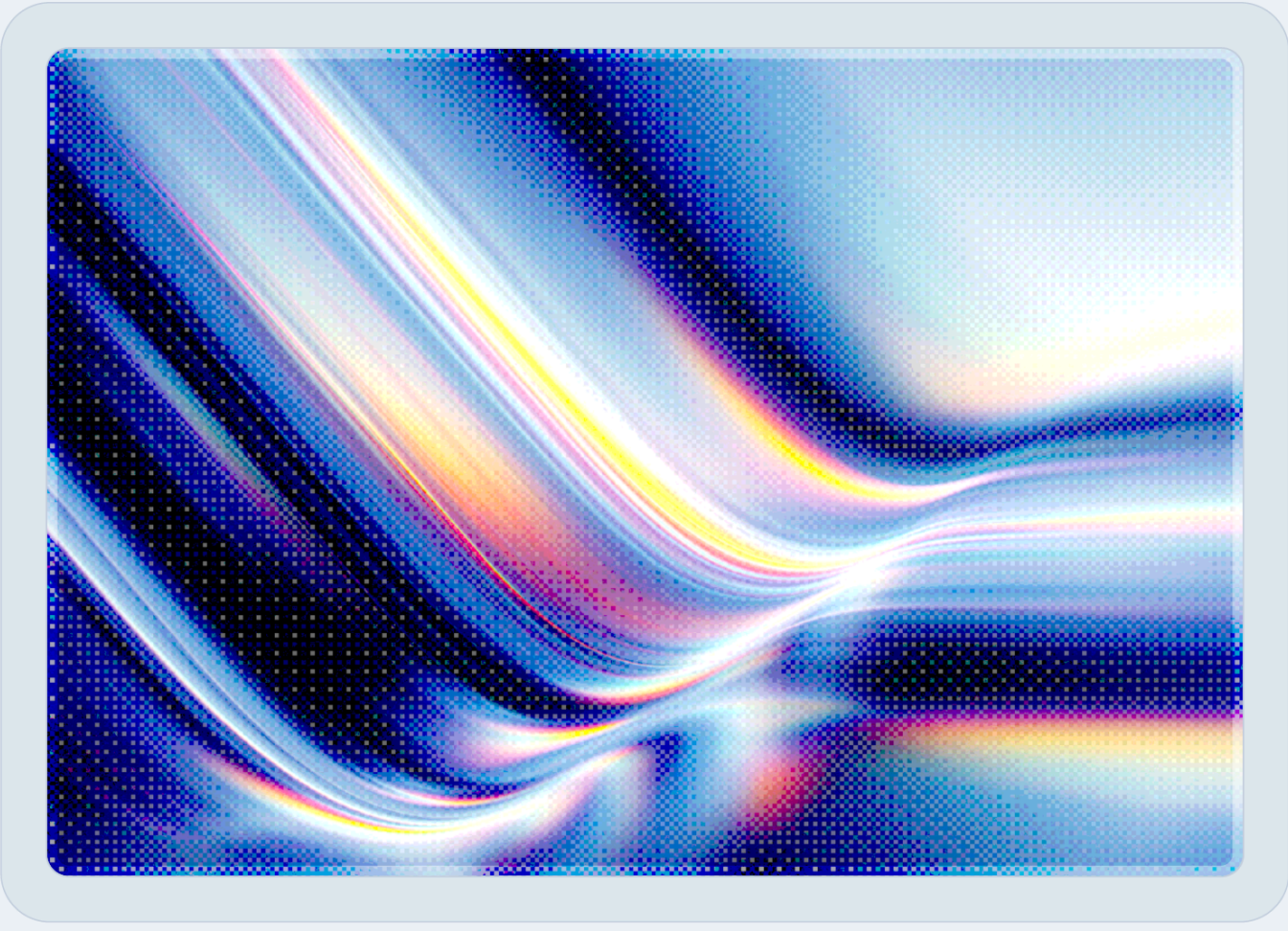
CSAI Foundation | Cloud Security Alliance

CISA OT Zero Trust: AI Governance for Industrial Systems

Federal Guidance Addresses Zero Trust Architecture and AI Integration Risks in Operational Technology Environments

2026-05-01

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Key Takeaways

- On April 29, 2026, CISA and federal partners published "Adapting Zero Trust Principles to Operational Technology," a 28-page guide organized around the six functions of the NIST Cybersecurity Framework 2.0 and designed to help critical infrastructure operators overcome the unique constraints of industrial environments [1][2].
- A companion document, "Principles for the Secure Integration of Artificial Intelligence in Operational Technology," co-authored with Australia's Australian Cyber Security Centre and six other national cybersecurity agencies and published in December 2025, establishes four governance principles specifically governing AI deployments within OT environments [3].
- The guidance acknowledges that standard IT security controls can be ineffective and potentially dangerous in OT settings, where safety-critical operations demand continuous availability over intrusive security enforcement [1].
- The AI-OT companion guidance identifies model drift, data poisoning, and LLM hallucination as acute risks in industrial contexts and states that large language models "should almost certainly not be used to make safety decisions" autonomously [3].
- In December 2025, the Department of Defense published complementary guidance specifying 105 zero-trust activity requirements across seven pillars for OT environments, with target completion tied to fiscal year 2027 for IT systems and an open timeline for OT [5].
- The issuance of these two related federal documents—the December 2025 AI-OT guidance and the April 2026 zero trust guide—suggests that regulators are treating AI integration as an inseparable dimension of industrial cybersecurity posture, not a future consideration.

Background

Operational technology encompasses the programmable systems that interact directly with physical processes: industrial control systems (ICS), SCADA platforms, programmable logic controllers (PLCs), distributed control systems, and the sensors and actuators that govern power grids, water treatment facilities, oil and gas pipelines, and manufacturing lines. Unlike IT environments—where failures are typically recoverable, though often at significant cost—OT failures can halt physical processes, damage

equipment, and endanger human life. This safety imperative has historically justified the isolation of OT networks from corporate IT and the public internet—the so-called "air gap" that practitioners and regulators long treated as the primary line of defense.

That perimeter has eroded significantly over the last decade. The industry-wide push toward Industry 4.0, remote monitoring, and predictive maintenance has accelerated IT-OT network convergence [1], exposing industrial systems to threat vectors that were once theoretically distant. Adversaries have responded. Nation-state groups such as Volt Typhoon have specifically targeted OT environments to establish persistent footholds, and malware families such as CrashOverride and BlackEnergy have demonstrated the capability to interfere with physical processes through software manipulation alone [1]. Industry surveys of operational technology infrastructure consistently document systemic exposure: significant proportions of industrial control systems run outdated operating systems, lack basic endpoint protection, and maintain direct internet connections that contradict the air-gap assumptions underlying their original security designs [1][4]. Against this backdrop, CISA's new guidance represents a formal acknowledgment that perimeter-based thinking is insufficient and that zero trust principles must now extend into industrial environments.

The April 2026 guidance reflects a second inflection point that has emerged alongside the zero trust imperative: the active integration of artificial intelligence into OT operations. Utilities and manufacturers are deploying AI-driven predictive maintenance systems, AI-powered anomaly detection platforms, and increasingly autonomous process optimization agents that interact directly with industrial control systems. Each of these deployments introduces new attack surfaces—adversarial manipulation of training data, model drift under operational conditions, and the possibility that AI-driven decisions propagate into safety-critical control loops faster than human operators can intervene. CISA's December 2025 companion guidance on AI in OT addresses this dimension directly, and together these two federal documents represent the most operationally specific guidance to date on treating AI governance as a structural component of industrial cybersecurity [3].

Security Analysis

The CISA OT Zero Trust Guide: Structure and Core Findings

The April 29 guide, "Adapting Zero Trust Principles to Operational Technology," is a joint publication by CISA, the Department of Energy, the Federal Bureau of Investigation, the Department of State, and the Department of Defense [2]. Its structure maps directly onto the six core functions of NIST CSF 2.0—Govern, Identify, Protect, Detect, Respond, and Recover—adapted for the specific constraints of

industrial environments. The document aligns with CISA's Cross-Sector Cybersecurity Performance Goals 2.0, the DoD Zero Trust Reference Architecture v2.0, NIST SP 800-82 Revision 3, and the internationally recognized ISA/IEC 62443 series [1].

The guide's most consequential contribution is its treatment of the tension between zero trust's core requirement of continuous verification and OT's requirement for uninterrupted real-time operations. Standard zero trust tooling—network access brokers, continuous authentication agents, endpoint detection and response software—can introduce latency, require system reboots, or demand network reconfiguration in ways that are operationally unacceptable and potentially unsafe on plant floors. The guidance does not resolve this tension by prescribing one approach, but instead frames a risk-based, phased implementation model that prioritizes passive monitoring and compensating controls where active enforcement is not feasible. Network TAPs and SPAN ports, for example, can support asset inventory and behavioral baselining without generating any traffic that could disrupt a control loop [1][4].

The guide introduces a zones-and-conduits model as the organizing framework for OT network segmentation, building on ISA/IEC 62443 concepts. Security zones are defined groupings of OT assets with similar security requirements and communication profiles; conduits are the controlled, policy-enforced pathways between zones. Supply chain risk management receives explicit attention as a high-priority area, with the guidance calling for procurement requirements that mandate logging capabilities, identity management support, and secure-by-default communication protocols in newly purchased OT components [1]. The table below summarizes the guide's treatment of each NIST CSF 2.0 function in the OT context.

NIST CSF 2.0 Function	OT-Specific Emphasis
Govern	Establish cross-functional ownership between cybersecurity and OT operations teams; define acceptable risk thresholds for industrial processes
Identify	Build and continuously maintain real-time asset inventories using passive monitoring; classify devices by criticality and communication dependencies
Protect	Apply least-privilege access with MFA where technically feasible; segment networks into zones and conduits; address supply chain identity and logging requirements
Detect	Deploy OT-protocol-aware monitoring tools; adapt behavioral baselines to industrial process patterns rather than IT network traffic norms

NIST CSF 2.0 Function	OT-Specific Emphasis
Respond	Develop OT-specific incident response playbooks that account for safety interlocks, manual override procedures, and regulatory notification requirements
Recover	Establish offline backup and business continuity procedures that can restore operations without reconnecting to potentially compromised network segments

Secure AI Integration in OT: The Companion Guidance

Published in December 2025—five months before the April 2026 zero trust guide—the companion document, "Principles for the Secure Integration of Artificial Intelligence in Operational Technology," was co-authored by CISA and Australia's Australian Signals Directorate's Australian Cyber Security Centre, with participation from the NSA Artificial Intelligence Security Center, the FBI, and the national cybersecurity agencies of Canada, Germany, the Netherlands, New Zealand, and the United Kingdom [3]. Its publication signals that federal and allied government bodies now treat AI deployment in OT as a governance problem requiring explicit principles, not merely a set of technical controls to be retrofitted to existing frameworks.

The guidance establishes four core principles. The first, "Understand AI," addresses education and literacy: organizations must ensure that OT operators, engineers, and security personnel understand AI system behavior, failure modes, and the ways AI-driven outputs can differ qualitatively from outputs produced by deterministic rule-based systems. The second, "Assess AI Use in OT," requires organizations to conduct structured business-case evaluations before deploying AI in industrial settings, explicitly weighing the operational benefits of predictive maintenance or anomaly detection against the risks introduced by model dependencies and data pipeline exposure. The third principle, "Establish Governance and Assurance," calls for testing AI models in simulated OT environments before production deployment, maintaining AI-specific risk registers separate from traditional IT/ICS controls, and demanding that vendors provide Software Bills of Materials (SBOMs) that enumerate AI model components alongside software dependencies. The fourth principle, "Embed Safety and Security Practices," requires that human operators retain meaningful override capability, that AI systems include fail-safe fallback mechanisms, and that architectural separation keep AI processing infrastructure off plant floors where feasible, with only sanitized data flowing between operational networks and AI platforms [3][8].

The guidance's treatment of large language models is notably cautious. It states explicitly that LLMs can fabricate plausible but false outputs and that such systems should almost certainly not be used to make safety decisions in OT environments [3]. This reflects a wider concern among industrial security practitioners that the opacity of LLM reasoning—and the potential for prompt injection or adversarial input manipulation—makes them unsuitable for direct integration into control loops, even as they find legitimate application in adjacent use cases such as maintenance documentation retrieval and operator training interfaces.

AI-Specific Risks in Safety-Critical OT Environments

The AI-OT guidance identifies a cluster of risks that are qualitatively distinct from those already documented for AI in enterprise IT environments. Model drift—the gradual degradation of a model's accuracy as real-world operating conditions diverge from the distribution of its training data—is particularly hazardous in industrial settings because the degradation may be silent, manifesting as incrementally less reliable anomaly detection or predictive maintenance alerts rather than an obvious system failure [3]. An OT security platform whose underlying model has drifted may continue to report confidence scores that appear normal while its actual detection capability has eroded substantially.

Data poisoning poses a complementary threat. OT environments generate highly structured time-series data from sensors and actuators; an adversary with access to that data pipeline—whether through a compromised historian server, a manipulated maintenance interface, or a corrupted software update to a field device—could inject subtly anomalous readings designed to bias an AI model's learned baseline of normal operations [3][9]. The guidance further highlights the risk that operational data may persist statistically within trained model weights beyond formal data retention periods, creating potential privacy and security exposures when models are shared with vendors for retraining or when model files are transferred across organizational boundaries without adequate access controls [3]. For organizations deploying AI-powered security agents in OT environments, the guidance warns that a compromised or manipulated agent can become an attacker-controlled pathway rather than a detection capability—a concern that directly parallels the emerging threat class of adversarial manipulation of AI-based security tooling [3].

Recommendations

Immediate Actions

Organizations operating critical infrastructure OT environments should prioritize passive asset discovery as the foundational step before any zero trust implementation, using network TAPs or SPAN ports to build accurate inventories of all devices, communication patterns, and protocol usage without introducing active scanning traffic that could disrupt control loops. The CISA guide identifies asset visibility as a foundational prerequisite for zero trust implementation [1]. Concurrently, security and operations teams should conduct an AI audit of all deployed or planned AI systems in the OT environment, inventorying their data inputs, outputs, update mechanisms, and connections to control system networks. Any AI system that can directly influence a setpoint, interlock, or actuator without mandatory human confirmation should be flagged for immediate architectural review against the AI-OT companion guidance criteria [3].

Short-Term Mitigations

Over the next three to six months, organizations should develop or update their OT-specific incident response playbooks to incorporate AI system failure scenarios explicitly, including procedures for detecting model drift, disabling a compromised AI agent while maintaining process continuity, and restoring operations from a known-clean model checkpoint. Supply chain risk management deserves parallel attention: procurement teams should begin requiring that OT vendors provide SBOMs inclusive of AI components and contractually commit to notification windows for model updates that may affect anomaly detection baselines [1][3]. Network segmentation efforts should use the zones-and-conduits model to isolate AI processing infrastructure from plant-floor control networks, ensuring that only sanitized data flows into AI platforms and that AI-generated outputs feed into human-reviewed dashboards rather than directly into control system parameters [3][10].

Strategic Considerations

The convergence of zero trust architecture and AI integration in OT signals a deeper organizational challenge: the security team, the OT engineering team, and (for organizations using AI-enabled industrial platforms) the data science team must develop genuine cross-functional fluency if governance frameworks are to shape actual risk posture rather than produce documentation without operational effect. The CISA guidance calls explicitly for institutionalized collaboration between these groups at the policy level, not merely coordination during incidents [1]. Organizations that are beginning zero trust

maturity assessments should also treat AI governance maturity as a parallel dimension to be evaluated, using the four principles of the companion guidance as a scoring rubric: Where does the organization stand on AI literacy among OT personnel? Is there a structured process for assessing AI business cases against OT risk tolerances? Are AI-specific risk registers in place and actively maintained? Do deployed AI systems have documented fail-safe mechanisms and human override procedures? Answering these questions honestly before expanding AI capabilities in industrial environments will determine whether the guidance shapes actual risk posture or merely generates documentation.

CSA Resource Alignment

The CISA OT zero trust and AI governance guidance connects directly to several CSA research areas and frameworks. CSA's "Zero Trust Guidance for Critical Infrastructure" (2024) provides a 64-page technical implementation guide that operationalizes zero trust for OT and ICS environments using CSA's five-step Zero Trust implementation process, including detailed coverage of the Purdue Model, ISA/IEC 62443 zones and conduits, and OT-specific policy enforcement point design—making it a directly applicable companion resource for organizations implementing the new CISA guidance [6]. CSA's "Zero Trust Guidance for Achieving Operational Resilience" extends this foundation to cover the interplay between zero trust architecture and business continuity, a dimension the CISA guide treats but does not develop in depth [7].

From an AI governance perspective, CSA's MAESTRO framework (Multi-Agent Extensible Security Taxonomy and Risk Observatory) provides threat modeling constructs for agentic AI systems that map directly onto the CISA companion guidance's concerns about AI agents as high-value targets and the risk of compromised agents becoming attacker-controlled pathways. MAESTRO's Layer 4 (Deployment and Infrastructure) and Layer 5 (Monitoring and Observability) are particularly applicable to the OT context, where infrastructure constraints limit the deployment options available for AI monitoring tooling. CSA's AI Controls Matrix (AICM), as a superset of the Cloud Controls Matrix (CCM), provides a controls catalog that can be used to assess AI governance posture against the four principles in the CISA AI-OT guidance. The "Confronting Shadow Access Risks" guidance, which addresses Zero Trust for AI and LLM deployments, is also directly applicable where organizations are deploying LLM-based interfaces against OT historian or maintenance documentation systems—a use case that the CISA companion guidance flags as requiring careful access control design.

References

[1] CISA. "Adapting Zero Trust Principles to Operational Technology." CISA, April 29, 2026.

[2] CISA. "CISA and U.S. Government Partners Unveil Guide to Accelerate Zero Trust Adoption in Operational Technology." CISA, April 29, 2026.

[3] CISA and ASD ACSC. "Principles for the Secure Integration of Artificial Intelligence in Operational Technology." CISA, December 2025.

[4] Industrial Cyber. "New CISA Guidance Outlines Zero Trust Roadmap for OT Environments Facing Legacy Constraints and Growing Attack Surfaces." Industrial Cyber, April 2026.

[5] DefenseScoop. "Pentagon Posts Guidance on Implementing Zero Trust for OT." DefenseScoop, December 1, 2025.

[6] Cloud Security Alliance. "Zero Trust Guidance for Critical Infrastructure." CSA, 2024.

[7] Cloud Security Alliance. "Zero Trust Guidance for Achieving Operational Resilience." CSA, 2024.

[8] CISA. "New Joint Guide Advances Secure Integration of Artificial Intelligence in Operational Technology." CISA, December 2025.

[9] SecureWorld. "NSA/CISA Guidance Demands a Secure-by-Design Approach for AI in OT." SecureWorld, December 2025.

[10] Industrial Defender. "CISA and Partners Release Principles for Secure Integration of AI in OT." Industrial Defender, December 2025.

