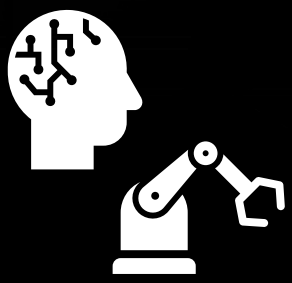
Palindrome Technologies

ASSURANCE | TRUST | CONFIDENCE



Can AI Be Trusted in Industrial Automation?

The Case for Security Assurance and Validation

PALINDROME TECHNOLOGIES
www.palindrometech.com

Contents

1 Introduction 3
2 Enter the Era of Operational AI 3
3 The Extended Attack Surface Created by Industrial AI 5
4 Why Traditional Security Testing Is Not Enough 8
5 Establishing Trust Through Security Validation 10
6 Human Oversight, Governance, and Accountability 13
7 Security Assurance as a Lifecycle Engineering Discipline 15
8 Building Trustworthy Autonomous Operations 16
9 Conclusion 19
10 References 21

Figures

Figure 1 Deterministic automation and AI-enabled industrial automation comparison. 4
Figure 2 industrial AI attack surface examples and operator decision pathways. 6
Figure 3 Traditional OT cybersecurity testing and AI security validation as complementary assurance activities 9
Figure 4 Industrial AI Trust Validation Framework 11
Figure 5 Risk-based human oversight model 14
Figure 6 Lifecycle assurance checkpoints for industrial AI security validation. 15
Figure 7 Evidence-based progression toward trustworthy autonomous operations. 19

Abstract

Artificial Intelligence is becoming embedded in industrial automation systems that monitor assets, optimize processes, support predictive maintenance, prioritize alarms, and increasingly influence operational decisions. This shift changes the basis for trust because AI-enabled functions depend on data quality, model behavior, deployment context, supplier components, operator interpretation, and operating conditions that may change after deployment. Traditional cybersecurity controls remain necessary for protecting industrial automation and control systems, but they are not sufficient to determine whether an AI-enabled function will remain reliable, explainable for operational use, resilient to manipulation, and bounded by appropriate human oversight. This paper argues that industrial AI assurance must be treated as a lifecycle engineering discipline that connects ISA/IEC 62443-based industrial cybersecurity architecture, ISO/IEC 42001-based AI governance, deployed-system assurance, independent testing, and continuous monitoring. It introduces a practical trust-validation framework organized around architecture and control verification, operational behavior validation, adversarial testing, and deployment assurance with continuous monitoring. The paper further explains how AI expands the industrial attack surface across data pipelines, model repositories, inference services, AI platforms, supplier update paths, and operator decision pathways, and why validation must examine both conventional cybersecurity compromise and AI-specific failure modes such as corrupted data, model drift, manipulated inputs, automation bias, and misplaced reliance on uncertain recommendations. By interconnecting security validation, governance, human oversight, and evidence-based commissioning, the paper provides a structured approach for suppliers, integrators, and asset owners to establish confidence before AI-enabled products and autonomous functions are allowed to influence industrial operations.

Keywords: Industrial AI, Industrial Automation, AI Security Assurance, AI Security Validation, IACS Cybersecurity, ISA/IEC 62443, ISO/IEC 42001, AI Risk Management, Operational Technology Security, Adversarial Machine Learning, Model Drift, Data Integrity, Human Oversight, Deployed-System Assurance, Independent Testing, Trustworthy Autonomous Operations.

1 Introduction

For decades, industrial automation has been built on deterministic logic, where controllers execute defined instructions and operators supervise processes through specific procedures and alarm-management practices, making system behavior easier to specify, test, document, and repeat because defined system conditions are expected to produce consistent outputs. The advent of Artificial Intelligence shifts this paradigm because AI-enabled systems are often non-deterministic in practice, with outputs that may vary based on training data, model state, probabilistic inference, externally supplied logic or analytics accessed through APIs, changing context, and the quality of human operational inputs.

AI/ML alters the security assurance expectations for industrial systems because its capabilities and performance depend on data quality, model assumptions, training history, deployment context, and operating conditions that may evolve after deployment, which means a model that performs well during development can degrade in the field, respond incorrectly to corrupted inputs, or produce recommendations that appear plausible but are unsafe in a specific operational context. This non-deterministic behavior does not make AI unsuitable for industrial automation, but it underscores that trust and security assurance must be implemented differently from current approaches. Trustworthy AI requires stronger evidence of model accuracy and enhanced traditional cybersecurity controls against adversarial attacks, because industrial organizations need evidence that AI systems and the products that incorporate them remain secure, reliable, sufficiently explainable for operational use, resilient under adverse conditions, and bounded by human oversight and governance. The practical question is not simply whether AI can improve industrial performance, but whether organizations can demonstrate security assurance and risk management on a continuing basis.

This paper argues that security validation must become a continuous engineering discipline for industrial AI and for the products that incorporate AI-enabled capabilities, rather than a separate effort outside the existing industrial cybersecurity body of practice. The ISA/IEC 62443 series provides the cybersecurity engineering structure for industrial automation and control system (IACS) security, ISO/IEC 42001 provides the management-system structure for governing AI risk, roles, responsibilities, monitoring, and continual improvement, and ISASecure Automation and Control System Security Assurance (ACSSA) offers a structured path for evaluating deployed control systems and related security programs against selected ISA/IEC 62443 requirements. Independent testing strengthens this model by providing objective evidence that products, systems, and deployed environments have been evaluated against recognized criteria rather than only internal assumptions, vendor claims, or limited pilot results.

2 Enter the Era of Operational AI

Industrial automation has historically relied on deterministic control logic, safety functions, and human-supervised decision processes in which system behavior is easier to specify and test because defined inputs, operating states, and control instructions are expected to produce repeatable outputs. AI changes the basis for operational confidence by introducing systems that infer patterns, recommend actions, and increasingly influence the timing and priority of

operational decisions across manufacturing, energy, telecommunications, healthcare, utilities, transportation, and other critical infrastructure sectors, where AI is becoming embedded in systems that monitor assets, forecast failures, optimize process conditions, and coordinate complex operations. Figure 1 illustrates this shift by contrasting the predictable flow of deterministic automation with the data-driven and probabilistic behavior of AI-enabled automation.

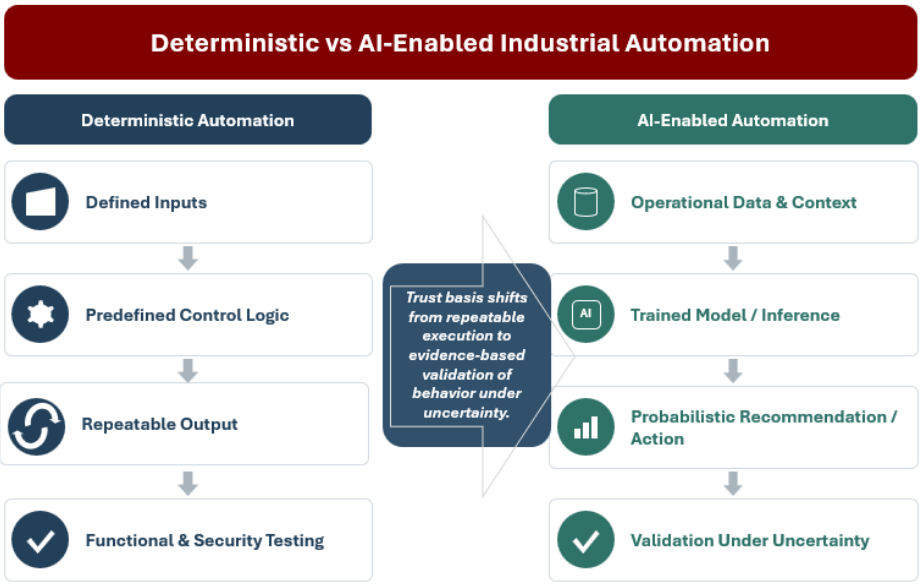


Figure 1 Deterministic automation and AI-enabled industrial automation comparison.

The comparison highlights a shift in the epistemic basis for operational trust in which conventional automation can be evaluated through specification conformance, repeatability, and control-path verification, whereas AI-enabled automation requires evidence that system outputs remain valid, interpretable for operational use, and bounded as data quality, model behavior, and operating context change. This distinction matters because AI can detect early signs of equipment degradation, identify process inefficiencies, support condition-based maintenance, and help operators interpret large volumes of operational data, while the same capabilities increasingly place AI within the decision fabric that shapes maintenance schedules, production adjustments, alarm prioritization, and control recommendations. As a result, confidence can no longer rest on pilot results or model performance claims alone and must instead be supported by repeatable validation, operational monitoring, and clear rules for when human review or manual intervention is required. The drivers that explain why AI changes the basis for operational trust include:

- **Determinism**: deterministic automation is expected to produce repeatable outputs under defined conditions, while AI-enabled automation depends on data, models, context, and probabilistic inference.
- **Operational Influence**: AI is moving from advisory analysis toward functions that shape maintenance timing, process optimization, alarm prioritization, and operational situational awareness.

- **Risk Consequence**: incorrect AI outputs become more significant when they can affect safety, production continuity, quality, compliance, or critical infrastructure functions.
- **Evidence-Based Trust**: confidence must be established through validation, monitoring, and governance rather than assumed from pilot results, model accuracy, or supplier claims.

The practical implication is that AI changes how industrial organizations must define, test, monitor, and govern confidence before allowing model-generated outputs to influence operational decisions.

3 The Extended Attack Surface Created by Industrial AI

Traditional cybersecurity programs protect industrial automation assets (e.g., PLCs, DCSs, HMIs, engineering workstations, remote access paths, and OT networks) from common attack vectors, including unauthorized access, lateral movement, or protocol manipulation of control communications. AI-enabled automation compounds the attack surface by adding less visible dependencies that influence operational decisions, including operational data pipelines, training environments, model repositories, inference engines, edge or cloud deployment platforms, supplier-maintained update paths, and third-party software components. The challenge therefore expands from protecting control assets against direct compromise to preserving the integrity of the information, models, platforms, and human decision pathways that determine how those assets are operated. These challenges are further compounded by adversaries' use of AI to develop new attack vectors and increase attack velocity.

Figure 2 provides examples of the expanded assurance boundary by connecting traditional OT assets to the dimensions that now shape industrial AI risk, including unauthorized access, data integrity, runtime input integrity, model integrity, AI platforms, supply-chain dependencies, and operator decision pathways. These dimensions represent a connected attack surface because a weakness in one area can change the behavior, interpretation, or operational consequence of another.

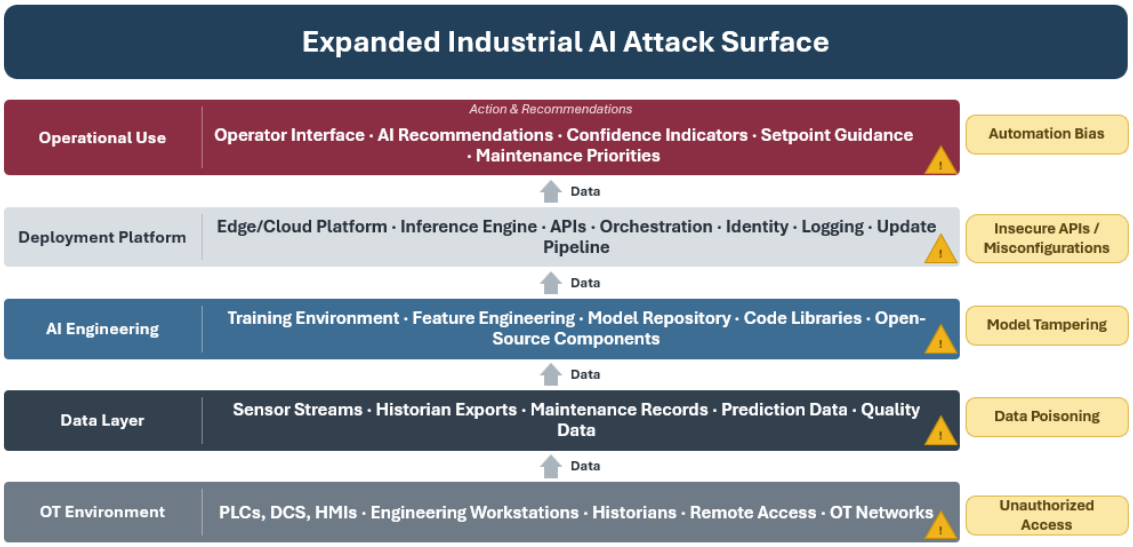


Figure 2 industrial AI attack surface examples and operator decision pathways.

For industrial AI, the assurance boundary must extend across the data, decision, and control chain that links field data, historian records, model development, deployment infrastructure, supplier updates, operator interpretation, and operational action. ISA/IEC 62443 provides a useful foundation for this analysis because its treatment of zones, conduits, security levels, roles, and technical requirements can be applied to the new trust boundaries created by AI components. The resulting threat model is broader than direct controller compromise because operational consequences can arise through unauthorized access, corrupted data, manipulated inputs, compromised models, platform weaknesses, supplier dependencies, or misplaced reliance on AI-generated outputs.

Several risk dimensions follow from this expanded assurance boundary and illustrate how AI-enabled automation can be manipulated, degraded, or misapplied without always presenting as a conventional OT intrusion:

- **Unauthorized access:** compromised credentials, exposed engineering workstations, misconfigured remote access, or supplier connectivity can provide an entry point for manipulating AI-enabled automation. In AI-enabled automation, adversaries may use this access not only to reach a controller, but also to influence the data, model, configuration, or recommendation pathways that shape operational decisions. Unauthorized access can also be used to alter sensor streams, modify model artifacts, change inference configurations, corrupt training records, or influence operator-facing recommendations.
- **Data integrity compromise:** industrial AI derives its operational judgment from historical records, sensor values, maintenance data, and historian exports. If these sources are incomplete, stale, falsified, or deliberately manipulated, the model may adopt an inaccurate definition of normal behavior or produce recommendations that do not reflect the actual process condition. A predictive maintenance system trained on corrupted

equipment-performance data may therefore miss early degradation or recommend actions that increase rather than reduce operational risk.

- **Runtime input manipulation:** this risk extends data-integrity concerns into live operation because false telemetry, spoofed sensor values, incomplete data streams, or manipulated environmental readings can influence AI outputs without changing the model itself. In an autonomous or semi-autonomous setting, the resulting recommendation may affect maintenance timing, process setpoints, quality decisions, alarm prioritization, or resource allocation, making input integrity a direct operational assurance concern rather than only a data-management issue.
- **Model, platform, and supply-chain compromise:** industrial AI often relies on open-source frameworks, commercial platforms, cloud services, pre-trained models, external datasets, vendor-managed updates, and deployment pipelines. Weaknesses in these components can affect downstream industrial systems even when local OT segmentation and access controls appear sound, which means product assurance must examine not only the deployed model but also the platforms, repositories, update mechanisms, and supplier obligations that sustain it.
- **Operator decision-support risk:** AI recommendations influence how operators interpret alarms, prioritize maintenance, respond to abnormal conditions, and decide whether to intervene. Automation bias, the tendency to over-rely on automated recommendations when they appear credible or have been useful in the past, is one expression of this risk. As recommendations prove useful over time, operators may become less inclined to challenge them, creating excessive trust that becomes dangerous when model performance drifts, operating conditions depart from the training environment, or adversarial manipulation causes the system to produce confident but unreliable outputs.

These risk dimensions show why the expanded industrial AI attack surface must be addressed as an engineering assurance problem rather than only as a network-security problem. Several implications follow for AI-enabled IACS design and operation:

- AI expands the attack surface from control systems and OT networks to data pipelines, model repositories, inference services, AI platforms, supplier update paths, and operator decision interfaces.
- AI introduces new trust boundaries around data flows, model repositories, inference services, and supplier-maintained components that must be represented in the IACS architecture rather than treated as external analytics infrastructure.
- Operational harm can originate from corrupted information, compromised models, manipulated inference pathways, supplier-introduced weaknesses, or misplaced reliance on AI-generated recommendations even when a PLC or control server has not been directly compromised.

- The security architecture must account for how access control, data integrity, model integrity, platform assurance, supplier governance, and operator-facing information interact when AI-generated outputs influence operational decisions.

In practice, AI components should be treated as operationally significant elements of the IACS and assigned explicit trust boundaries, data-quality expectations, update controls, monitoring requirements, and operator-facing confidence information. Once AI-generated outputs influence maintenance timing, process optimization, alarm prioritization, or control recommendations, the supporting data pipelines, model repositories, inference services, supplier dependencies, and operator interfaces become part of the industrial cybersecurity architecture and must be engineered, monitored, and governed accordingly.

4 Why Traditional Security Testing Is Not Enough

Traditional security testing is effective at identifying vulnerabilities in industrial assets and networks, but AI-enabled automation introduces assurance concerns that depend on data validity, model behavior, inference context, and operator interpretation rather than on exploitability alone. ISA/IEC 62443 provides the risk-based structure for incorporating AI components into the IACS security architecture by linking risk assessment, zones and conduits, target security levels, secure development practices, and system security requirements to the control-allocation process that governs conventional IACS functions.

This structure is necessary because the test objective expands from determining whether an adversary can gain access, manipulate configuration, or interfere with IACS assets to determining whether the AI-enabled function remains operationally valid when the quality, timing, completeness, or representativeness of its inputs changes. A platform may satisfy relevant cybersecurity controls and still produce unsafe or misleading recommendations if model behavior drifts, confidence information is incomplete, or the operating state falls outside the validated conditions established during development and validation.

Figure 3 positions traditional OT cybersecurity testing and AI security validation as complementary assurance activities that address different failure mechanisms in connected industrial systems.

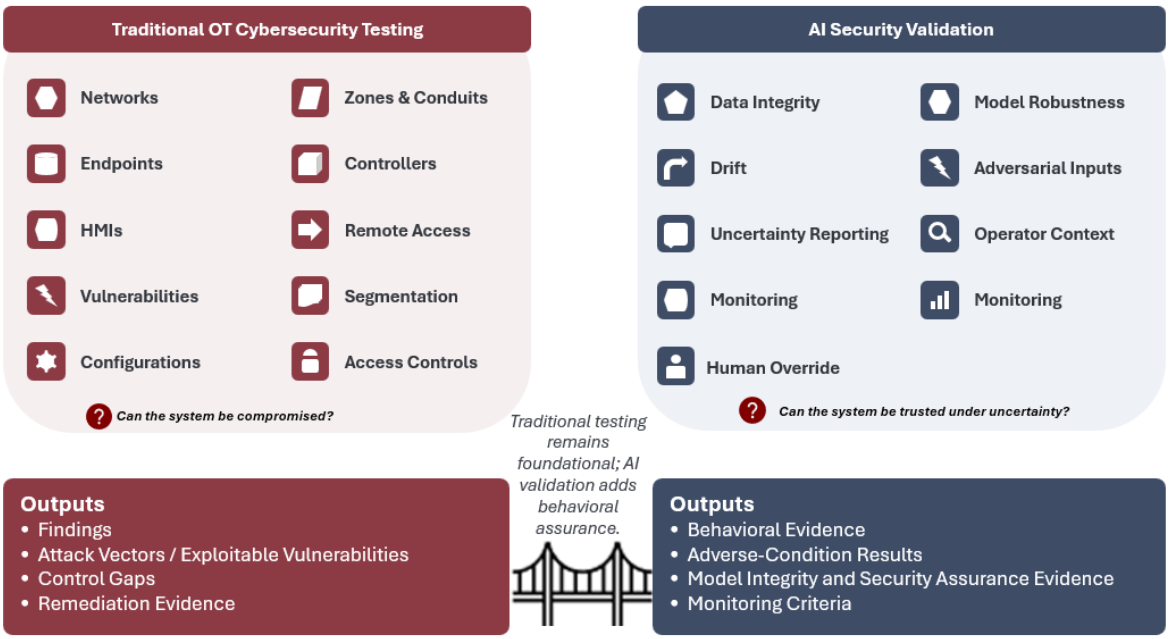


Figure 3 Traditional OT cybersecurity testing and AI security validation as complementary assurance activities

The comparison separates two assurance questions that are often conflated in connected industrial systems. Traditional OT testing asks whether an adversary can gain access, move laterally, change configuration, or interfere with control assets, while AI validation asks whether the AI-enabled function remains dependable when inputs, assumptions, operating modes, or process conditions depart from the expected range. This distinction is important for automation environments because an unsafe recommendation may arise from corrupted data, model drift, or misplaced operator reliance without appearing as a conventional intrusion, yet the effect may still be visible in production quality, equipment reliability, safety margin, or continuity of service.

Independent testing becomes important when supplier evidence and internal engineering review do not fully resolve the risk of connecting an AI-enabled product to an operating environment. A neutral assessment can examine security functions, configuration assumptions, update mechanisms, model behavior, operator-facing information, and documentation as deployment conditions rather than as isolated product features, giving procurement, engineering, cybersecurity, and operations teams a stronger basis for acceptance decisions.

AI assurance must therefore evaluate whether the function remains dependable under incomplete data, abnormal operating states, adversarial inputs, uncertainty, and changing environmental conditions, while also determining whether operators receive sufficient context to interpret confidence, limitations, and consequence before acting on a recommendation. For industrial products that perform monitoring, optimization, remote access, analytics, or advisory functions, this evaluation should occur before site integration because unreliable behavior discovered after procurement may already be embedded in configuration choices, operating procedures, and user expectations.

The validation scope should therefore include operational questions that reflect how AI-enabled functions fail in industrial use, including:

- How does the function respond when sensor values, historian records, maintenance data, or process context are incomplete, delayed, stale, or corrupted?
- Can recommendations be shifted through manipulated inputs, adversarial manipulation, or changes in data distribution that do not trigger conventional security alarms?
- Does performance remain acceptable when operating modes, equipment condition, product mix, load profile, or environmental conditions change?
- Are confidence levels, assumptions, data limitations, and operational consequences visible in a form that supports informed operator judgment?
- Can operators intervene, override, escalate, or place the function in a safe, degraded, or manual mode when outputs appear unsafe, uncertain, or inconsistent with process knowledge?

These questions convert AI validation into an operational assurance activity because they test whether the function remains reliable, bounded, and operationally defensible when plant data, process conditions, engineering assumptions, and operator expectations diverge from the controlled environment in which the capability was developed or demonstrated. For asset owners, this evidence reduces uncertainty before procurement acceptance and integration, while for suppliers it creates a disciplined path for improving product maturity and substantiating security and dependability claims against expectations that customers increasingly treat as part of industrial product quality.

5 Establishing Trust Through Security Validation

Security validation for industrial AI should be organized around the evidence needed to trust an AI-enabled function in its intended IACS environment. The relevant evidence begins with the industrial cybersecurity architecture, including risk assumptions, trust boundaries, zones and conduits, target security levels, supplier responsibilities, service-provider obligations, secure development practices, and technical controls [1]. It also includes AI governance evidence, including defined accountability, risk treatment, change control, monitoring, documentation, performance evaluation, and continual improvement for AI-enabled functions [3]. Palindrome's proposed trust-validation framework translates these concepts into four practical assurance checkpoints: architecture and control verification, operational behavior validation, adversarial testing, and deployment assurance with continuous monitoring.

Figure 4 presents this proposed framework as a pragmatic testing and assurance process for establishing trust before AI-enabled outputs influence industrial operations.

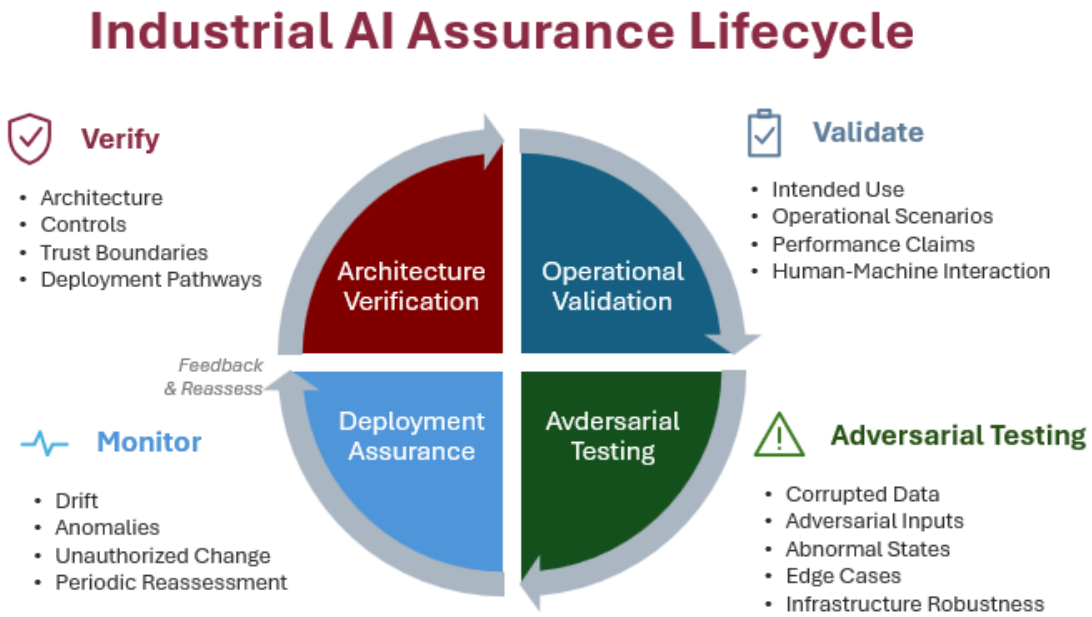


Figure 4 Industrial AI Trust Validation Framework

The framework is intended to prevent AI deployment from being justified by a single model metric, pilot demonstration, or supplier assertion. Each assurance checkpoint produces evidence that answers a different assurance question: whether the function is placed correctly within the IACS architecture, whether its behavior is reliable under representative operating conditions, whether it can be manipulated or degraded through adversarial activity, and whether the deployed environment continues to support the accepted risk posture over time. This approach allows product security, AI behavior, integration assumptions, asset-owner procedures, and monitoring evidence to be evaluated as parts of one trust case rather than as separate reviews.

Architecture and control verification confirms that the AI-enabled function is represented within the industrial security architecture before it is trusted in operation. This assurance checkpoint reviews trust boundaries, data flows, identity and access controls, segmentation, remote access paths, model repositories, deployment pipelines, logging, update mechanisms, supplier dependencies, and service-provider interfaces. The purpose is to determine whether the product or function can be integrated, configured, monitored, and constrained in a manner consistent with the intended IACS risk posture and applicable system and component security requirements [1].

Operational behavior validation determines whether the AI-enabled function performs reliably for the industrial task it is expected to support, such as anomaly detection, predictive maintenance, process optimization, alarm prioritization, remote advisory support, or autonomous adjustment. This assurance checkpoint evaluates accuracy, consistency, decision quality, uncertainty communication, explainability for operational use, and operator interpretation under representative process conditions. The validation threshold should be tied to the consequence

of the function because advisory analytics, supervisory recommendations, and autonomous actions require different levels of evidence.

Adversarial testing evaluates whether the AI-enabled function can be manipulated, misled, or degraded in ways that may not appear as conventional OT intrusions. Test conditions should include corrupted or stale datasets, manipulated sensor values, abnormal process states, communication disruptions, load or product-mix changes, unauthorized configuration changes, compromised model artifacts, and adversarial manipulation of inputs or prompts where applicable. Independent adversarial testing is valuable because it examines the product or system from an external, security-informed, and operationally realistic perspective, revealing weaknesses that may remain hidden when testing is limited to expected use cases, nominal process behavior, or supplier-defined demonstrations [8], [9].

Deployment assurance and continuous monitoring sustains the trust case after integration by tracking model drift, data-quality degradation, anomalous behavior, unauthorized change, configuration drift, software updates, supplier dependencies, and shifts in operator reliance against defined operating baselines. This assurance checkpoint also reviews asset-owner procedures, service-provider practices, risk assessment evidence, technical security configuration, incident response, reassessment intervals, and authority to override or disable the function. Deployed-system assurance is important because a model that was reliable at acceptance may become less dependable as the plant, process, threat landscape, supplier ecosystem, or automation architecture changes [2].

The framework can be operationalized through four evidence-producing assurance checkpoints:

- **Architecture and control verification:** confirm that trust boundaries, data flows, access controls, model repositories, deployment pathways, supplier-controlled components, logging, and update mechanisms are represented in the IACS architecture and supported by evidence.
- **Operational behavior validation:** determine whether the AI-enabled function performs reliably under representative operating conditions and whether claims about accuracy, consistency, explainability, uncertainty communication, and operational usefulness are supported by evidence.
- **Adversarial testing:** expose the function to manipulated, degraded, abnormal, and adversarial conditions to determine whether behavior remains constrained and operationally acceptable when assumptions fail or inputs are intentionally distorted.
- **Deployment assurance and continuous monitoring:** track drift, anomalies, configuration changes, updates, data quality, supplier dependencies, and operator reliance after deployment, while periodically reassessing whether the operating environment continues to support the trust case.

This process converts AI security validation from a model-review exercise into an industrial assurance discipline that connects product design, system integration, operational behavior, adversarial resistance, governance evidence, and deployed-system monitoring.

6 Human Oversight, Governance, and Accountability

Human oversight in AI-enabled automation should be treated as part of the control strategy rather than as an administrative safeguard added after deployment. The relevant engineering question is where human judgment must remain in the operating loop when AI influences alarms, maintenance timing, production adjustments, environmental controls, safety margins, or service-continuity decisions. Oversight therefore depends on the consequence of the function, the reliability of the AI output, the time available for intervention, and the operator's ability to understand why a recommendation was produced.

Operator trust is an instrumentation problem as much as a governance problem because the operator needs usable visibility into the data sources, confidence level, assumptions, operating limits, and consequence of acting on an AI-generated output. Without that information, the human role can degrade from informed supervision to confirmation of a model's recommendation, which is particularly problematic where decisions affect safety, production continuity, environmental performance, regulatory obligations, or critical infrastructure services.

Governance becomes effective when it is expressed through operating procedures, responsibility assignments, change-control rules, acceptance criteria, and escalation paths that engineers and operators can apply during normal and abnormal conditions. An asset-owner security program can be extended to cover model ownership, validation responsibility, data stewardship, risk acceptance, monitoring, incident response, periodic reassessment, and authority to override or disable an AI-enabled function [1]. AI management-system practices strengthen this governance layer by requiring defined roles, risk treatment, documentation, performance evaluation, monitoring, and continual improvement for AI-enabled functions [3]. In this role, governance defines how autonomy is commissioned, constrained, supervised, modified, and withdrawn over time.

Oversight should scale with the function's operational consequence and intervention window. Low-impact optimization or advisory functions may be governed through periodic review and performance monitoring, medium-consequence functions may require operator confirmation, alarm-based escalation, or threshold limits, and high-consequence functions should include pre-defined approval workflows, manual override, fallback modes, stop conditions, and independent validation evidence before autonomous action is permitted. Figure 5 introduces the escalation model by linking operational consequence to the degree of review, confirmation, override capability, fallback design, and independent validation evidence required before AI-generated outputs are allowed to influence operations.

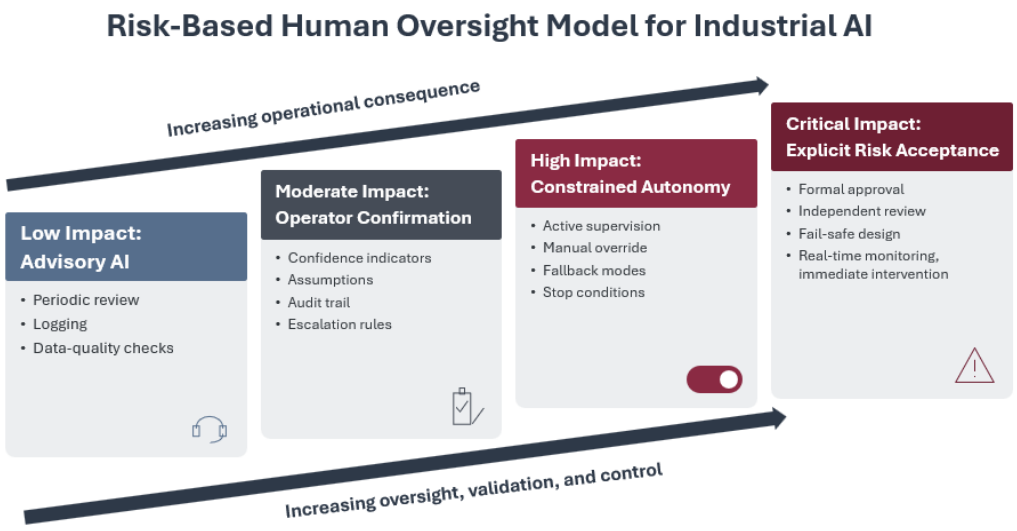


Figure 5 Risk-based human oversight model

The escalation model treats oversight as control allocation across advisory, supervisory, and autonomous operating modes. Its value is that it links the required level of human review and technical fallback to the consequence of an incorrect output, rather than assuming a uniform oversight requirement for every AI function. Low-impact advisory functions can often be governed through periodic review, while functions that may affect safety, reliability, compliance, production continuity, or critical infrastructure services require intervention paths, override authority, fallback behavior, and independent validation commensurate with their operational consequence. Several design principles follow for AI-enabled operational decision-making including:

- Oversight intensity should be tied to operational consequence, the operator intervention window, and the ability to place the function in a safe, degraded, or manual mode.
- Operator interfaces should expose the data basis, confidence level, assumptions, operating limits, and consequence of acting on the recommendation.
- Governance should define model ownership, validation responsibility, AI management-system responsibilities, risk acceptance, change control, incident response, reassessment intervals, and authority to override, suspend, or withdraw the function.
- Autonomous functions should include escalation logic, manual override, fallback modes, abnormal-output investigation, and evidence thresholds for expanding operational authority.

The engineering objective is to keep autonomy observable, interruptible, and accountable at the points where model-generated outputs can affect the process, while allowing lower-consequence applications to scale with controls proportionate to their operational effect.

These oversight and governance requirements become durable only when they are embedded into the lifecycle of design, integration, commissioning, operation, change management, and reassessment.

7 Security Assurance as a Lifecycle Engineering Discipline

Industrial AI assurance becomes a lifecycle engineering discipline because the trust case is shaped before deployment by architecture choices, data selection, model design, supplier components, integration assumptions, operator workflows, and monitoring obligations. The validation framework introduced in Section 5 provides a practical way to manage that lifecycle by treating architecture and control verification, operational behavior validation, adversarial testing, and deployment assurance as recurring assurance checkpoints rather than as one-time acceptance activities.

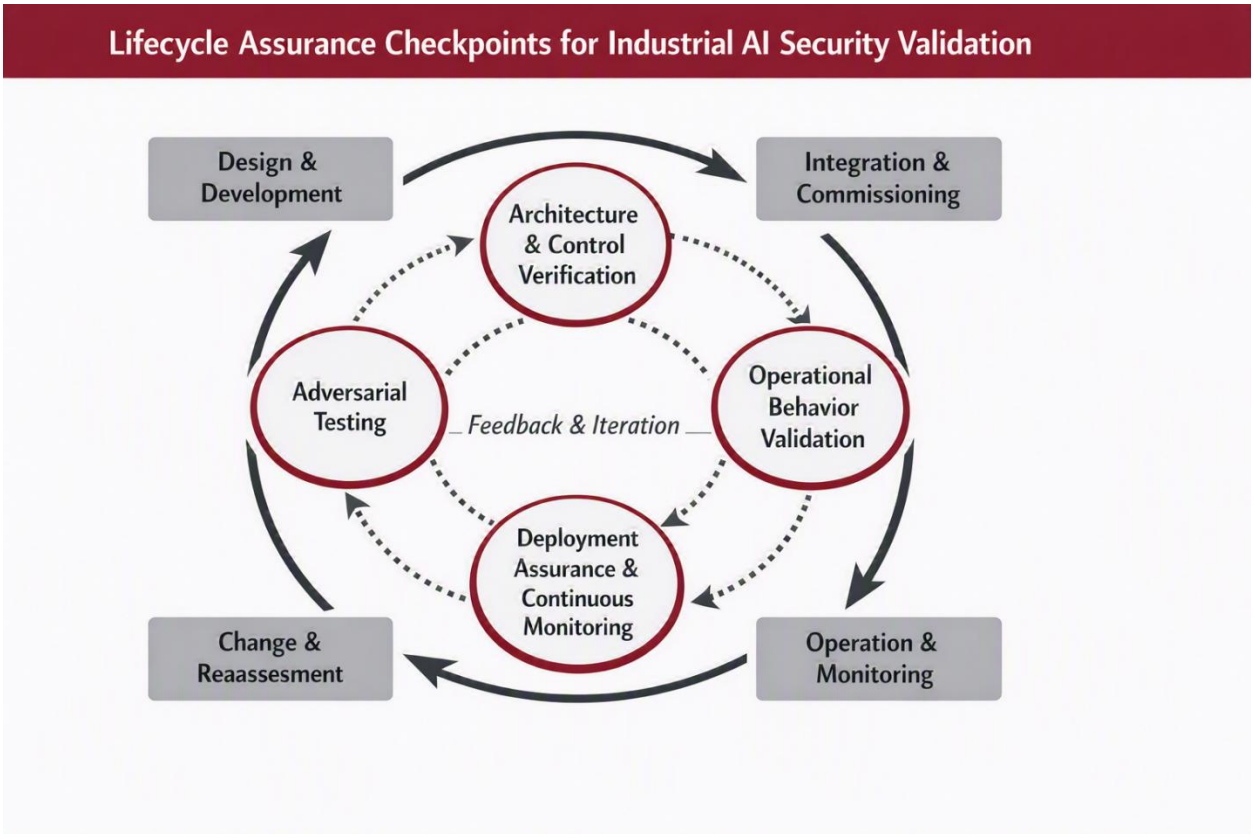


Figure 6 Lifecycle assurance checkpoints for industrial AI security validation.

During design and development, the first assurance checkpoint requires the intended operational use to be translated into trust boundaries, data provenance requirements, access controls, supplier obligations, update mechanisms, monitoring points, oversight assumptions, and acceptance evidence. These decisions determine whether the AI-enabled function can later be validated against the intended IACS environment rather than assessed only against laboratory performance, vendor demonstrations, or isolated model metrics.

The second assurance checkpoint examines operational behavior under representative process conditions. Validation should determine whether the function's outputs remain reliable, explainable for operational use, and interpretable by operators when equipment condition, product mix, load profile, environmental conditions, data quality, or operating mode changes. This moves validation beyond model accuracy and into the practical question of whether the function can support an industrial decision without creating uncertainty that operators cannot recognize or manage.

The third assurance checkpoint applies adversarial testing to determine whether the function can be manipulated, misled, or degraded through corrupted data, manipulated inputs, abnormal process states, unauthorized configuration changes, compromised model artifacts, supplier-introduced changes, or adversarial prompts where applicable. This checkpoint is important because AI-enabled failures may not appear as conventional OT intrusions even when they produce operational consequences.

The fourth assurance checkpoint extends assurance into deployment by monitoring whether the trust case remains valid as the plant, process, threat landscape, supplier ecosystem, and automation architecture change. Operating evidence should include model performance, data quality, configuration changes, software updates, supplier dependencies, operator reliance, incident response, reassessment intervals, and authority to override, suspend, or withdraw the function. Deployed-system assurance is especially important because a reliable acceptance result does not guarantee continued reliability after integration into a changing industrial environment [2].

Recent research on LLM-assisted IEC 62443 risk assessment reinforces the need for this evidence discipline because AI-generated assessment artifacts can vary in threat-scenario structure, zoning granularity, target security level assignment, and architectural coherence. These artifacts can support early engineering analysis, but they should be treated as decision-support outputs that require traceability, assumption management, independent review, and expert validation before they are used to justify security requirements or architecture decisions [10].

When applied in this way, lifecycle assurance converts AI security from a point-in-time acceptance decision into a continuing engineering discipline that connects supplier evidence, system integration, operational behavior, adversarial resistance, governance evidence, change management, and monitoring.

Independent testing strengthens this discipline by aligning supplier claims with the evidence needed for procurement, acceptance, integration, commissioning, and lifecycle maintenance decisions. In this role, testing helps convert cybersecurity and AI dependability from post-development review items into design attributes that affect product maturity, deployment risk, and customer confidence. Once this lifecycle evidence discipline is established, it becomes the basis for deciding whether an AI-enabled function should remain advisory, progress to supervisory influence, or be permitted to take autonomous action.

8 Building Trustworthy Autonomous Operations

Autonomous operation should be treated as a staged transfer of operational authority rather than as a feature that is enabled once the underlying AI capability appears technically mature. For industrial organizations, the governing question is what evidence justifies allowing an AI-enabled function to move from advisory output to supervisory influence or autonomous action within defined operating conditions, especially when the function can affect safety margins, continuity, quality, regulatory compliance, environmental controls, or critical infrastructure service delivery. Governance evidence should show that accountability, risk treatment, monitoring, change control, and continual-improvement processes are in place before authority is expanded [3].

A risk-based adoption model builds confidence by controlling the function's authority over the process. Advisory functions can be observed against operating data without direct control influence, supervisory functions can require operator confirmation or threshold-based escalation, and autonomous functions should be permitted only when validation, fallback behavior, override authority, monitoring, and residual-risk acceptance are aligned with the consequence of the action.

Resilience in autonomous operations should be engineered as constrained behavior under degraded or abnormal conditions. Defined operating conditions, abnormal-condition detection, uncertainty communication, graceful degradation, safe-state behavior, manual intervention paths, and escalation logic define how the function behaves when data, equipment state, process conditions, communications, or supplier dependencies depart from the assumptions used during development and validation.

Independent validation strengthens the trust case by testing whether the AI-enabled function remains dependable under the validated operating conditions where it will be used, while giving suppliers, integrators, and service providers a disciplined way to substantiate secure development, service delivery, product security, and deployment claims. Deployed-system assurance evaluates whether the operating IACS, asset-owner program, service-provider practices, risk assessment process, and technical controls continue to support the intended risk posture after the product is integrated into the plant or infrastructure environment [2].

For autonomous operations and AI-enabled industrial products, assurance should be translated into commissioning, procurement, and change-management criteria that determine when an AI-enabled function may progress from advisory output to supervisory influence or autonomous action. These criteria should connect the operating mode, authorized decision scope, process consequence, validation evidence, interlock and fallback design, and residual-risk acceptance rather than treat AI deployment as a feature-enablement decision:

- Commission AI-enabled functions by operating mode, beginning with advisory output and expanding toward supervisory influence or autonomous action only after evidence shows that the function remains dependable within its defined process conditions, authorized decision scope, and operator intervention window.
- Map AI components into the IACS architecture so data pipelines, model repositories, inference services, supplier-maintained components, and operator interfaces have

explicit trust boundaries, target security levels, control allocations, and ownership responsibilities [1].

- Scale validation depth to process consequence so functions that can affect safety margins, production continuity, product quality, regulatory compliance, environmental controls, or critical infrastructure service delivery require stronger evidence before authority is expanded beyond advisory use.
- Engineer autonomous behavior within defined operating conditions, abnormal-condition detection, uncertainty indication, authorization conditions, interlocks where appropriate, manual override, fallback modes, safe-state behavior, and escalation logic before the function is permitted to shape process action without direct confirmation.
- Use independent testing, AI-specific validation, and deployed-system evidence to connect supplier claims, product security behavior, integration assumptions, validated operating limits, service-provider dependencies, and the asset owner's intended risk posture [2].

A measured adoption path treats autonomy as progression through defined operating modes, authorized decision scope, and intervention requirements. Risk-based architecture, target security levels, and control allocation establish where the AI-enabled function belongs in the IACS environment [1]. AI governance evidence establishes accountability, risk treatment, monitoring, and continual improvement before authority is expanded [3]. Deployed-system assurance determines whether the operating environment continues to support the intended risk posture [2]. AI-specific validation and independent testing address model behavior, product security claims, supplier dependencies, validated operating limits, and assumptions that conventional control-system testing was not designed to evaluate.

Figure 7 organizes the evidence required before an AI-enabled function is allowed to assume greater operational authority. The figure emphasizes that authority should expand only when the trust case combines architecture and control evidence, governance evidence, product and supplier evidence, operational behavior evidence, adversarial testing evidence, deployed-system evidence, and human oversight.

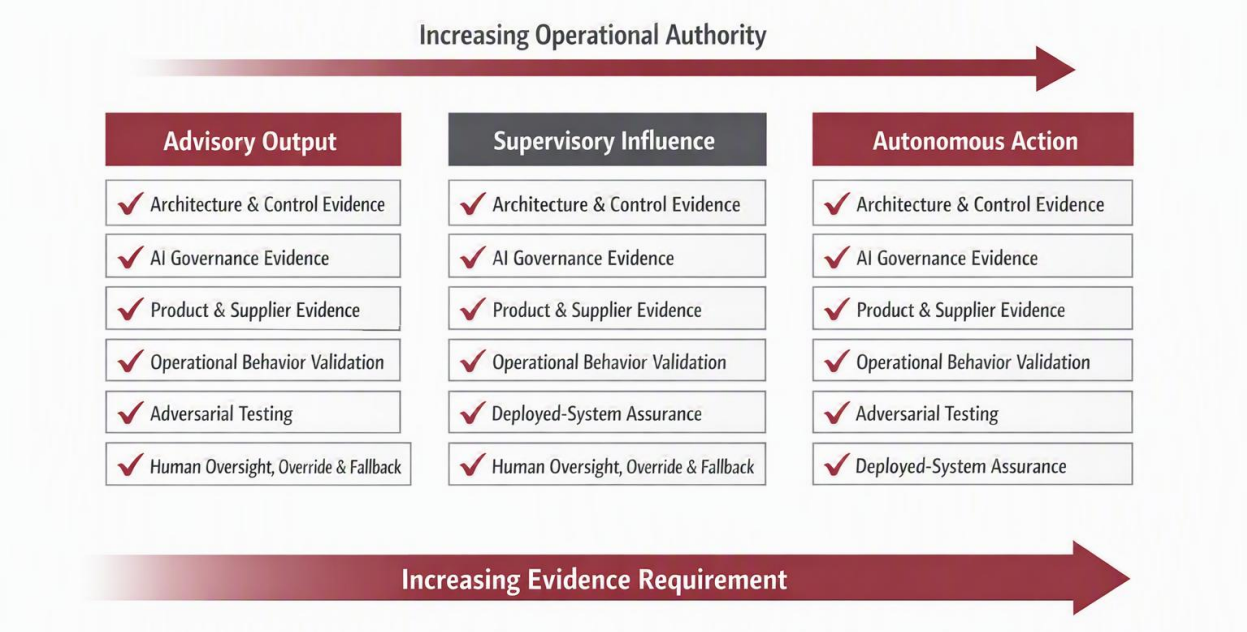


Figure 7 Evidence-based progression toward trustworthy autonomous operations.

The evidence-based progression should be read as a staged reduction of uncertainty before operational authority is expanded. Architecture and control evidence establishes where the AI-enabled function belongs in the IACS environment; AI governance evidence establishes accountability, risk treatment, monitoring, documentation, and continual improvement; product and supplier evidence substantiates security and dependability claims; operational behavior validation examines whether the function remains reliable within intended operating conditions; adversarial testing examines whether the function can be manipulated, degraded, or misled; deployed-system assurance determines whether the operating environment continues to support the intended risk posture; and human oversight constrains authority where operational consequence remains significant. The value of this progression is that it prevents autonomy from being justified by a single test result, model metric, or product claim and instead requires converging evidence across product behavior, system integration, deployed operation, governance, and human accountability.

9 Conclusion

AI is becoming a foundational element of modern industrial automation because it can improve maintenance, optimize operations, support digital twins, and enable more adaptive forms of operational decision-making. At the same time, AI changes the basis for operational trust. Confidence can no longer rest only on deterministic logic, conventional control-system testing, pilot demonstrations, or supplier claims, because AI-enabled functions depend on data quality, model behavior, deployment context, supplier components, operator interpretation, and operating conditions that may change after deployment.

The central assurance challenge is therefore not whether AI can be useful in industrial automation, but whether organizations can produce sufficient evidence that AI-enabled products

and functions remain secure, reliable, bounded, and operationally defensible in their intended IACS environments. Traditional cybersecurity remains essential, but it must be extended to account for AI-specific failure modes such as corrupted data, manipulated inputs, model drift, compromised model artifacts, adversarial conditions, automation bias, and misplaced reliance on uncertain recommendations.

This paper has argued that industrial AI assurance should be treated as a lifecycle engineering discipline built around recurring assurance checkpoints. Architecture and control verification establishes where the AI-enabled function belongs within the IACS security architecture; operational behavior validation determines whether the function performs reliably under representative process conditions; adversarial testing examines whether the function can be manipulated, degraded, or misled; and deployment assurance with continuous monitoring sustains the trust case as the plant, process, supplier ecosystem, threat landscape, and operating assumptions change. ISA/IEC 62443, ISO/IEC 42001, deployed-system assurance, independent testing, and risk-based human oversight provide complementary structures for producing and maintaining that evidence.

For autonomous industrial operations, this evidence must also govern how operational authority is expanded. Advisory output, supervisory influence, and autonomous action should be treated as distinct operating modes, each requiring evidence commensurate with the consequence of the function, the operator intervention window, the availability of override and fallback mechanisms, and the asset owner's residual-risk acceptance. Autonomy should therefore be commissioned through an evidence-based progression rather than enabled as a product feature once the underlying AI capability appears technically mature.

The future of trustworthy industrial AI will not be defined by model sophistication alone. It will be defined by the ability of suppliers, integrators, and asset owners to build and maintain a defensible trust case that connects security architecture, AI governance, product assurance, operational validation, adversarial testing, deployed-system monitoring, and human accountability. Organizations that make this evidence discipline part of procurement, commissioning, change management, and operations will be better positioned to realize the benefits of AI while preserving the safety, resilience, and confidence on which industrial automation depends.

10 References

[1] International Society of Automation, “ISA/IEC 62443 Series of Standards,” *ISA*. URL: https://www.isa.org/standards-and-publications/isa-standards/isa-iec-62443-series-of-standards

[2] ISASecure, “Automation and Control System Security Assurance (ACSSA) Certification,” *ISASecure*. URL: https://isasecure.org/acssa-certification

[3] International Organization for Standardization and International Electrotechnical Commission, “ISO/IEC 42001:2023 Information technology, Artificial intelligence, Management system,” ISO/IEC, 2023.

[4] Palindrome Technologies, “Managing AI Risk the Smart Way: Why ISO/IEC 42001 Can Be a Game Changer,” *Palindrome Technologies*, Apr. 13, 2025. URL: https://palindrometech.com/emerging-technologies-security-blog/managing-ai-risk-smart-way-why-iso/iec-42001-can-be-a-game-changer

[5] Palindrome Technologies, “Unlocking Industrial Cybersecurity: A Deep Dive into ISA/IEC 62443 Standards,” *Palindrome Technologies*, Jun. 9, 2025. URL: https://palindrometech.com/emerging-technologies-security-blog/unlocking-industrial-cybersecurity-a-deep-dive-into-isa/iec-62443-standards

[6] Palindrome Technologies, “Private 5G Security Uncovered: A Case Study (2024),” *Palindrome Technologies*, Mar. 30, 2025. URL: https://palindrometech.com/case-studies-listing-page/private-5g-uncovered-case-study

[7] P. Slattery, A. K. Saeri, E. A. C. Grundy, J. Graham, M. Noetel, R. Uuk, J. Dao, S. Pour, S. Casper, and N. Thompson, “The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks from Artificial Intelligence,” *arXiv*, arXiv:2408.12622, 2024. DOI: 10.48550/arXiv.2408.12622.

[8] National Institute of Standards and Technology, “Artificial Intelligence Risk Management Framework (AI RMF 1.0),” *NIST AI 100-1*, Jan. 2023. DOI: 10.6028/NIST.AI.100-1.

[9] A. Vassilev, A. Oprea, A. Fordyce, H. Anderson, X. Davies, and M. Hamin, “Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations,” *NIST AI 100-2e2025*, Mar. 2025. DOI: 10.6028/NIST.AI.100-2e2025.

[10] A. Bernhard and M. Pfister, “AI-Driven Risk Assessment for Critical Infrastructure Based on IEC 62443 Using Large Language Models,” *Research Square*, 2026. DOI: 10.21203/rs.3.rs-9050939/v1.

Contact Information

info@palindrometech.com

www.palindrometech.com

https://palindrometech.com/ai-cybersecurity

About Palindrome Technologies

Palindrome's organizational philosophy is built upon three fundamental principles

Assurance
Trust
Confidence

Founded in 2005, Palindrome Technologies Inc. is an applied cybersecurity research, testing, inspection, and certification organization focused on high-assurance technologies that support critical infrastructure, industrial automation, telecommunications, medical devices, and emerging AI-enabled systems. In industrial automation, Palindrome combines ISO/IEC 17025 laboratory testing, ISO/IEC 17065 product certification, ISA/IEC 62443 conformity assessment, ISASecure certification services, ACSSA-related deployed-system assurance, and ISO/IEC 42001 AI management-system certification to help suppliers and asset owners substantiate security, governance, and operational trust through objective evidence. Before forming Palindrome, the company's principals worked for Bellcore, formerly Bell Communications Research, in the Security and Fraud group, where they supported security assurance efforts for telecommunications providers, product vendors, and the U.S. government.

Since its inception, Palindrome has provided high-technology services related to securing emerging technologies, global enterprise organizations, including healthcare, financial services, energy, and government organizations, and carrier-grade networks. Palindrome is an accredited ISO/IEC 17025 testing laboratory as well as an FCC, GSMA, CTIA, and IEEE-designated cybersecurity testing laboratory. Palindrome has helped global enterprise organizations, service providers, and product vendors deploy and maintain secure networks, services, and products. The Palindrome team is also known for its contributions to industry standards bodies, including IEEE, GSMA, CTIA, and ATIS, and to branches of the U.S. government, including FCC CSRIC VII, CSRIC IX, CSRIC X and NIST.