MeruEPC
SECURE. SMART. HUMAN

I-CAIRA

The Integrated Cyber-AI Risk Assessment Framework

Pioneering Standards-Based Security for Industrial AI Systems

Definitive Edition | *Version 1.0*

November 2025

© 2025 MeruEPC Pte. Ltd.

Secure. Smart. Human.

Intellectual Property Notice

This framework and all derivative materials—including diagrams, risk matrices, control mappings, and methodologies—are the exclusive property of MeruEPC Pte. Ltd. and constitute proprietary intellectual property developed under the MeruEPC Cybersecurity Innovation Program.

All Rights Reserved. Unauthorized reproduction, distribution, or commercial use of this material without written consent from MeruEPC is strictly prohibited. This document may be registered with intellectual property authorities for additional protection.

Community Draft – Extended Public Consultation

Consultation Window: 10 November 2025 – 15 March 2026

Community Platform: I-CAIRA – Open Community for AI + OT Cyber Risk (LinkedIn)

Contact: vijay@meruepc.com

Join the Conversation: *This framework represents a pioneering effort to establish the definitive standard for AI security in operational technology. Your expertise—whether from asset ownership, system integration, research, or regulatory experience—will help shape the future of industrial AI governance. We welcome your feedback, real-world insights, and collaborative contributions.*

Join the Collaborative Effort Here: *https://www.linkedin.com/groups/16203032/*

Table of Contents

PART I: THE STRATEGIC IMPERATIVE

1. The Strategic Imperative for AI Risk Pioneering in OT

1.1 Defining the Convergence: AI, Cyber-Physical Systems, and IACS Criticality
1.2 The Pioneer Mandate: Aligning Responsible AI with Industrial Safety
1.3 Establishing the Governance Foundation: Integrating NIST CSF 2.0

2. AI/ML Deployment Context and Lifecycle Risks in OT

2.1 AI Use Cases and the Behavioral Baseline Challenge
2.2 The MLOps Security Lifecycle in OT
2.3 Cataloging Advanced Threats: The Adversarial AI Landscape

PART II: THE I-CAIRA METHODOLOGY

3. The I-CAIRA Methodology (Anchored to IEC 62443-3-2)

3.1 Harmonizing Standards: The I-CAIRA Workflow
3.2 Step 1 – Integrated Asset & AI Component Inventory
3.3 Step 2 – Consequence Analysis & Cyber-Physical Alignment
3.4 Step 3 – AI-Specific Likelihood & Threat Modeling
3.5 Step 4 – Quantified Risk Scoring for AI Systems

4. The MeruEPC AI Risk Assessment Standard Template

4.1 Template Structure
4.2 Integrated Risk Scenario Analysis
4.3 Mitigations & Control Recommendations

PART III: CORE FRAMEWORK

5. Purpose & Audience
6. Scope & Objectives
7. Key Definitions
8. Governance Model (NIST CSF 2.0 GV)
9. Lifecycle & Workstreams
10. Controls Catalogue (IEC 62443 mapped, AI-specific)
11. Assessment Method (6-Step Process)
12. Risk Scoring Rubric
13. Maturity Model (L0–L4)
14. Metrics, KRIs & SLOs
15. Implementation Roadmap (90-Day Starter)
16. Deliverables & Templates
17. Roles & RACI
18. Policies
19. Annexes
20. Comparative Landscape & Alignment
21. V1.1/V2.0 Recommendations

APPENDICES

Appendix A: Reference Standards & Frameworks
Appendix B: Control Mapping Tables
Appendix C: Example Risk Scenarios
Appendix D: Evidence Checklists
Appendix E: Community Consultation Timeline

About MeruEPC Pte. Ltd.

Contact Information & Community Participation.

Join the Collaborative Effort Here:

https://www.linkedin.com/groups/16203032/

PART I

THE STRATEGIC IMPERATIVE

1. The Strategic Imperative for AI Risk Pioneering in OT

The Industrial Automation and Control Systems (IACS) landscape is undergoing a critical transformation driven by the accelerated adoption of Artificial Intelligence and Machine Learning. These technologies are increasingly integrated into operational technology environments to enable anomaly detection, User and Entity Behavior Analytics (UEBA), predictive maintenance, and autonomous optimization. While these systems promise unprecedented efficiency gains, they introduce novel cyber-physical risks that traditional OT frameworks alone cannot fully address.

I-CAIRA emerges at this critical juncture as the pioneering framework that bridges AI governance with OT security, establishing MeruEPC as the industry leader in securing the next generation of intelligent industrial systems.

1.1 Defining the Convergence: AI, Cyber-Physical Systems, and IACS Criticality

Recent industry commitments to AI-powered Production System Optimization (AiPSO) in upstream operations and AI-integrated robotics for field services dramatically expand the attack surface—from the control room all the way to physical assets in the field. While IEC 62443 provides the risk-based backbone across the IACS lifecycle, AI's unique characteristics create new vulnerabilities:

- **Continuous Data Dependence:** AI systems require constant data feeds, creating persistent attack vectors through data poisoning and manipulation
- **Complex and Opaque Models:** Deep learning models operate as 'black boxes,' making it difficult to audit decisions or detect adversarial manipulation
- **Novel Attack Surfaces:** Model theft, adversarial inputs, and training data poisoning represent entirely new threat categories
- **Performance Drift:** Models degrade over time as operational conditions change, potentially leading to unsafe or insecure decisions

The Reality: Compliance checklists are insufficient. AI systems require continuous technical validation through MLOps practices embedded into the IEC 62443 lifecycle. I-CAIRA provides this integration.

1.2 The Pioneer Mandate: Aligning Responsible AI with Industrial Safety

What defines a pioneer in AI security?

Industry research identifies pioneers as organizations that act as *anticipators*—they don't wait for threats to materialize; they proactively build resilience into their systems. Pioneers invest 70% of their AI security effort into people and processes, not just technology. They place Responsible AI (RAI) principles at the core of their operations.

I-CAIRA positions MeruEPC and its clients as pioneers by:

- **Integrating Responsible AI with OT Safety:** Aligning AI/ML governance (MLOps) with OT security frameworks so that platforms like AiPSO and autonomous robotics are secured against adversarial threats and integrity failures from day one
- **Mandating Workforce Upskilling:** Ensuring teams can audit and secure intelligent systems, not just operate them
- **Addressing Cloud-Scale AI Risks:** As organizations form cloud-scale alliances for AI deployment, I-CAIRA provides the governance structure to manage high-scale, cloud-native AI risks
- **Enabling Continuous Adaptation:** Pioneers continuously adapt their risk monitoring and control processes—I-CAIRA's dynamic risk register and MLOps integration enable this

Table 1.1: Strategic Alignment – Mapping Trustworthy AI Principles to OT Imperatives

Trustworthy AI Principle	OT Cyber Implication	Direct I-CAIRA Assessment Focus
Technical Robustness & Safety	AI model must not induce process instability or unsafe states in physical systems	Adversarial resilience testing; continuous performance monitoring; safe-state enforcement
Transparency (Explainability)	Auditability of automated control and security decisions by operators and safety engineers	XAI documentation packages; output validation; decision rationale archives
Human Agency & Oversight	Operator retains override capability and ultimate accountability for system actions	Human-in-the-Loop (HIL) protocols; exception handling procedures; override logging
Privacy & Data Governance	Secure handling of OT telemetry and sensitive training data; protection against data exfiltration	Data poisoning prevention; strict access controls; lineage tracking; immutable audit logs
Fairness & Bias Mitigation	In OT, bias is a technical vulnerability: selection bias, structural bias, or validation gaps create unreliable predictions and unsafe actions	Treat bias as measurable operational risk; statistical validation; performance monitoring across conditions; forensic analysis when degradation detected

Critical Insight: *In OT environments, fairness isn't an ethical abstraction—it's a technical requirement. Bias creates operational vulnerabilities that can lead to equipment damage, safety incidents, or security breaches.*

1.3 Establishing the Governance Foundation: Integrating NIST CSF 2.0

NIST Cybersecurity Framework 2.0 introduces a critical sixth function: Govern (GV). This addition recognizes that effective cybersecurity starts with governance—the policies, roles, and oversight structures that guide an organization's risk management approach.

For AI in OT environments, the Govern function is essential because it:

- ✓ **Organizes AI Risk Strategy:** Defines acceptable AI use, risk appetite, and approval gates
- ✓ **Manages the AI Supply Chain:** Tracks model sources, training data provenance, and third-party dependencies
- ✓ **Establishes Roles and Accountability:** Clarifies who owns AI risks across OT Owner, AI Product Owner, MLOps Lead, Cybersecurity Lead, and Process Safety roles
- ✓ **Integrates with ERM:** Ensures AI risks are included in enterprise risk management frameworks

I-CAIRA adopts NIST CSF 2.0's Govern function as the organizational scaffold for all AI security activities, ensuring that technical controls are backed by clear governance structures. This integration distinguishes I-CAIRA from purely technical frameworks and positions MeruEPC as a strategic partner, not just a technical consultant.

2. AI/ML Deployment Context and Lifecycle Risks in OT

To secure AI effectively, we must understand how it's deployed, how it operates, and where vulnerabilities emerge throughout its lifecycle. This section examines the operational reality of AI in IACS environments.

2.1 AI Use Cases and the Behavioural Baseline Challenge

Common OT AI Applications:

- → **Anomaly Detection:** Identifying deviations from normal operational patterns to detect cyber threats, equipment failures, or process instability
- → **User and Entity Behaviour Analytics (UEBA):** Monitoring operator actions and system behaviours to identify insider threats or compromised credentials
- → **Predictive Maintenance:** Forecasting equipment failures before they occur, enabling proactive maintenance
- → **Production System Optimization (AiPSO):** Real-time optimization of industrial processes for efficiency, yield, or safety
- → **Autonomous Robotics:** AI-guided robots for inspection, maintenance, and field service operations

The Behavioural Baseline Challenge:

Many OT AI systems work by establishing a 'behavioural baseline'—a statistical model of what 'normal' looks like. The system then flags deviations from this baseline as potential threats or anomalies. This approach offers significant benefits:

- ✓ Rapid detection of previously unknown threats
- ✓ Reduced false positive rates compared to signature-based detection
- ✓ Adaptation to evolving operational conditions

However, this approach creates a critical vulnerability:

If an adversary can poison the training data during the baseline creation period, the system will learn to accept malicious activity as 'normal.' The compromised baseline effectively blinds the detection system to the very threats it's supposed to catch.

I-CAIRA Response:

- ✓ **Pre-Baselining Threat Assessment:** Mandatory security validation of the environment before baseline training begins
- ✓ **Forensic Data Validation:** Statistical analysis and expert review of training data to detect poisoning attempts
- ✓ **Baseline Verification:** Independent validation that the baseline represents genuinely normal behaviour
- ✓ **Continuous Baseline Monitoring:** Ongoing statistical tests to detect baseline drift or corruption

2.2 The MLOps Security Lifecycle in OT

Machine Learning Operations (MLOps) extends DevOps practices to AI/ML systems, covering the entire lifecycle from development through deployment and ongoing operation. In OT environments, MLOps must be secured at every stage to prevent compromise.

Phase 1: Development & Training

Objective: Build trustworthy models through rigorous data management and reproducible training processes

- ✓ **Data Integrity:** Cryptographic verification of training data sources; provenance tracking from sensor to model
- ✓ **Reproducibility:** Version control for datasets, code, hyperparameters, and dependencies; ability to recreate any historical model
- ✓ **Versioned Datasets:** Immutable storage with timestamps and digital signatures
- ✓ **Model Cards:** Comprehensive documentation of model purpose, training approach, performance metrics, limitations, and intended use cases

Phase 2: Deployment & Inference

Objective: Secure the production environment where models make real-time decisions affecting physical processes

- ✓ **Endpoint Authentication & Authorization:** Strong identity verification for inference API access; role-based access control
- ✓ **Secrets Management:** Secure storage and rotation of API keys, certificates, and credentials
- ✓ **Telemetry & Logging:** Comprehensive logging of inference requests, responses, and model decisions for audit and forensic analysis
- ✓ **Drift Detection:** Continuous monitoring for data drift, concept drift, and performance degradation

Phase 3: Conformity & Auditability

Objective: Demonstrate compliance with security standards and enable regulatory audits

- ✓ **Complete Lineage:** Traceable path from any model decision back through the model version, training data, and code used to create it
- ✓ **Audit Logs:** Tamper-evident logs of all model updates, configuration changes, and access events
- ✓ **Conformance Testing:** Automated validation that deployed models meet security requirements (adversarial robustness, explainability, performance)
- ✓ **External Validation:** Independent verification by third-party auditors or certification bodies

2.3 Cataloguing Advanced Threats: The Adversarial AI Landscape

AI systems face attack vectors that don't exist in traditional software. Understanding these threats is essential for effective risk assessment.

Threat Category 1: Data Poisoning

Attack Vector: Corruption of training data to induce specific model behaviors or degrade overall performance

OT Impact: Compromised anomaly detection (false negatives for attacks); biased predictive maintenance (missed failures); unsafe optimization recommendations

Detection Challenge: Sophisticated poisoning attacks are designed to be statistically subtle and difficult to distinguish from normal data variability

Threat Category 2: Adversarial Evasion

Attack Vector: Carefully crafted inputs designed to fool a trained model into making incorrect predictions

OT Examples:

• Manipulating sensor readings to evade intrusion detection while conducting cyberattacks

• Subtly altering process parameters to cause the AI to recommend unsafe operating conditions

Advanced Variant: Deep Reinforcement Learning (DRL) Evasion — Attackers use AI to learn how to evade detection systems. Research on the SWaT (Secure

Water Treatment) testbed demonstrated that DRL-based attackers could systematically identify and exploit blind spots in IDS systems

Threat Category 3: Model Theft

Attack Vector: Extraction of proprietary AI models through repeated queries to the inference API

OT Impact: Loss of competitive advantage from custom models; enables adversaries to conduct offline analysis to find weaknesses; intellectual property theft

Mitigation: Query rate limiting; API authentication; output randomization; monitoring for extraction patterns

Threat Category 4: Backdoor Attacks

Attack Vector: Insertion of hidden triggers into models during training that cause specific behaviours when activated

OT Scenario: A safety monitoring AI could be trained with a backdoor that causes it to ignore hazardous conditions when a specific, rare pattern of sensor readings occurs

Detection: Extremely difficult; requires careful examination of training data and model behaviour under diverse conditions

Critical Assessment Principle:

Treat unexplained performance degradation and emerging bias patterns as potential indicators of subtle adversarial compromise. When models begin to fail, investigate data lineage forensically—not just retrain blindly. Degradation may be an attack, not just drift.

PART II

THE I-CAIRA METHODOLOGY

3. The I-CAIRA Methodology (Anchored to IEC 62443-3-2)

I-CAIRA extends the IEC 62443-3-2 risk assessment methodology with AI-specific steps, creating a comprehensive approach that addresses both traditional OT risks and novel AI threats.

3.1 Harmonizing Standards: The I-CAIRA Workflow

The IEC 62443-3-2 standard defines a structured risk assessment process:

1. Define the System Under Consideration (SuC)
2. Partition the system into Zones and Conduits
3. Conduct Initial Risk Assessment
4. Conduct Detailed Risk Assessment and determine Security Level Targets (SL-T)

I-CAIRA inserts AI-specific assessment steps:

→ **AI Component Inventory:** Extended asset cataloguing to capture AI artifacts and dependencies
→ **Cyber-Physical Consequence Analysis:** Integration with ISA-TR84 and SCAI mapping to understand how AI failures affect safety systems
→ **AI-Specific Likelihood Modelling:** Threat assessment that accounts for adversarial ML attacks and model vulnerabilities
→ **Trustworthiness-Aware Risk Scoring:** Risk calculation that incorporates robustness, explainability, adversarial resilience, and bias

3.2 Step 1 – Integrated Asset & AI Component Inventory

The foundation of any risk assessment is knowing what you're assessing. For AI in OT, this means creating an integrated inventory that captures both traditional IACS assets and AI-specific components.

Extended Inventory Requirements:

For Each AI Component, Document:

✓ **Zone/Conduit Location:** Where the AI system resides in the IEC 62443 architecture
✓ **Functional Purpose:** What the AI system does (detection, prediction, optimization, control support)
✓ **Training Data Snapshot IDs:** Versioned identifiers for all training datasets
✓ **Feature Set:** Input variables and data sources the model uses

- ✓ **Model Algorithm & Version:** Type of AI/ML algorithm and specific version deployed
- ✓ **Hyperparameters:** Configuration settings that affect model behaviour
- ✓ **Performance Baselines:** Expected accuracy, precision, recall, and other metrics
- ✓ **XAI Documentation:** Explainability artifacts and decision rationale capabilities
- ✓ **Dependencies:** External data feeds, APIs, libraries, and services
- ✓ **Criticality Assessment:** Impact if the AI system fails or is compromised

Table 3.1: Integrated Asset & AI Inventory (Example Excerpt)

IACS Asset (Zone/Conduit)	AI/ML Functionality	Key Metadata Required	Potential AI Vulnerability	Criticality
Level 1: PLC/RTU (Zone 3 - Control)	Edge anomaly detection for process deviations	Training snapshot ID; performance metrics; XAI documentation; update frequency	Evasion attacks; model opacity; limited explainability at edge	**CRITICAL**
Level 3: Historian/SCADA Server (Zone 2 - Supervisory)	Predictive maintenance model for rotating equipment	Algorithm/version; feature store logs; validation results; retraining schedule	Training data poisoning; selection bias; concept drift	**HIGH**
Level 4: DMZ Gateway (Conduit 1 - IT/OT)	Network traffic anomaly detection and UEBA	Serving logs; monitoring alerts; API access controls; drift detection config	Model theft via API; secrets management; DRL evasion attacks	**CRITICAL**
SIS (Safety) (Zone 0 - Safety-Critical)	SCADA diagnostic ML for equipment health monitoring	Safety validation logs; SCAI mapping; performance under stress; HIL override capability	Performance bias leading to false trips or missed hazards; inadequate explainability for safety validation	**EXTREME**

Note: This inventory becomes the foundation for all subsequent risk analysis. Missing AI components from this inventory means missing risks in your assessment.

3.3 Step 2 – Consequence Analysis & Cyber-Physical Alignment

Traditional OT risk assessment focuses on single-point failures. AI systems, however, can fail in complex ways that affect multiple systems simultaneously. I-CAIRA extends consequence analysis to account for cyber-physical coupling and cascading failures.

Integration with ISA-TR84.00.09:

ISA Technical Report 84 provides guidance on safety-related control systems. I-CAIRA leverages this framework to map AI failure modes to physical consequences through Safety-Critical Application Integration (SCAI) analysis.

SCAI Mapping Process:

Step 1: Identify where AI systems interact with or influence safety-critical functions
Step 2: Analyze how AI failures (incorrect predictions, evasion, poisoning) could lead to hazardous states
Step 3: Consider coordinated cyber-physical scenarios beyond single-deviation HAZOP analysis
Step 4: Score consequences across multiple dimensions: Safety, Environmental, Production Loss, Regulatory

Multi-Dimensional Consequence Scoring

Score each potential consequence on a 1-5 scale across four dimensions:

Safety (S): 1 = Minor injury → 5 = Multiple fatalities

Environmental (E): 1 = Contained minor release → 5 = Major environmental disaster

Production (P): 1 = Minutes downtime → 5 = Extended shutdown, significant financial loss

Regulatory (R): 1 = Minor non-compliance → 5 = License suspension, criminal liability

Final Consequence Score = MAX(S, E, P, R) – Use the highest score to drive risk prioritization

3.4 Step 3 – AI-Specific Likelihood & Threat Modelling

Likelihood assessment must account for AI-specific attack vectors and vulnerabilities. Consider these factors:

- ✓ **Attacker Capability & Intent:** Sophistication required for adversarial ML attacks; motivation to target your industry
- ✓ **Exposure:** Remote management interfaces, IIoT connections, cloud dependencies
- ✓ **Model Class Weaknesses:** Known vulnerabilities in the AI algorithm type (e.g., deep learning susceptibility to adversarial examples)
- ✓ **Update Frequency:** How often models are retrained and deployed (more frequent = more attack opportunities)
- ✓ **Supplier Security Posture:** Security maturity of third-party AI vendors

3.5 Step 4 – Quantified Risk Scoring for AI Systems

I-CAIRA extends traditional L × C risk scoring with AI-specific modifiers that capture trustworthiness dimensions.

$$\mathbf{RISK = L \times C \times (1 + f(R, X, A, B) / k)}$$

Where:

L = Likelihood (1-5, AI-aware factors)

C = Consequence (1-5, maximum across dimensions)

R = Robustness score (0-5, lower is better)

X = Explainability score (0-5, higher risk if opaque)

A = Adversarial surface score (0-5, higher if exposed/weak)

B = Bias/performance variability score (0-5, higher if unstable)

k = Calibration constant (typically 4-20, organizationally tuned)

Risk Bands & SL-T Mapping:

1-5 (Low): SL-T 1 – Basic controls

6-12 (Moderate): SL-T 2 – Enhanced monitoring and drift detection

13-20 (High): SL-T 3 – Rigorous controls, continuous monitoring, mandatory XAI

>20 (Extreme): SL-T 4 – Maximum controls, real-time assurance, independent certification, HIL oversight

4. The MeruEPC AI Risk Assessment Standard Template

I-CAIRA provides a standardized deliverable format that ensures consistency, comprehensiveness, and auditability across AI risk assessments.

4.1 Template Structure

The I-CAIRA Assessment Report includes four core sections:

5. **Executive Summary & Risk Tolerance Statement**

High-level findings, risk posture, and recommendations for leadership

6. **Scope & AI Asset Inventory**

Complete catalog of AI components as per Table 3.1 format

7. **Integrated Risk Scenario Analysis**

Detailed risk scenarios with likelihood, consequence, and AI-refined scores

8. **Mitigation Plan & Control Mapping**

Specific controls mapped to IEC 62443 SRs and AI control families, with implementation timelines and evidence requirements

4.2 Integrated Risk Scenario Analysis (Example)

Table 4.1: Integrated Risk Scenario Analysis

Zone/Conduit	AI Component	Threat Scenario	Consequence	Likelihood	AI Risk Score	Required SL-T
Zone 3 (Control)	PID Calibration AI	DRL evasion attack learns to avoid detection	Process oscillation → equipment damage	Moderate (3)	**18 (High)**	**SL 3**
Conduit 1 (DMZ)	Central Anomaly Detection AI	Poisoned retraining dataset	False negatives → breach goes undetected	High (4)	**23 (Extreme)**	**SL 4**
SIS (Safety-Critical)	SCADA Diagnostic ML	Performance bias under stress conditions	False trip / missed hazard detection	Low (2)	**21 (Extreme)**	**SL 4**
Zone 1 (Upstream)	AiPSO Reservoir Optimization	Poisoning of geological/simulation data	Suboptimal production → financial loss + safety risk	Moderate (3)	**16 (High)**	**SL 3**
Field (Robotics)	Autonomous Inspection Robot	Navigation AI compromise → physical damage	Robot collision → equipment damage + injury risk	Low (2)	**12 (Moderate)**	**SL 2**

Note: These scenarios demonstrate how I-CAIRA addresses both traditional OT applications and emerging use cases like AiPSO and autonomous robotics—positioning MeruEPC at the forefront of industrial AI security.

PART III

CORE FRAMEWORK REFERENCE

5. Purpose & Audience

I-CAIRA is a practical, standards-aligned framework to govern, assess, and continuously assure Artificial Intelligence and Machine Learning used in Operational Technology and Industrial Automation & Control Systems.

Target Audience:

- Asset owners operating safety- and mission-critical environments
- System integrators implementing AI solutions in industrial settings
- Vendors developing AI-enabled OT products
- Regulators and auditors assessing AI governance

6. Scope & Objectives

Scope:

AI/ML models deployed anywhere in the OT stack: edge devices (PLC/RTU), SCADA/Historian systems, DMZ, SOC/SIEM that influence detection, diagnostics, prediction, optimization, control, or decision support.

Objectives:

- Make AI deployments safe, secure, and explainable across the full lifecycle
- Harmonize IEC 62443, NIST CSF 2.0 (Govern/GV), ISA TR84 safety with MLOps practices
- Provide repeatable assessment method, risk scoring, maturity model, and evidence templates

7. Key Definitions

AI Component: Model + data + code + runtime + configuration + lineage metadata

Model Risk: Risk that a model's failure or compromise leads to safety, environmental, financial, or compliance impact

Adversarial ML: Attacks targeting model/data (poisoning, evasion, model theft, backdoors)

XAI (Explainable AI): Techniques that make model behaviour interpretable and auditable

8. Governance Model (NIST CSF 2.0 GV)

Built on NIST CSF 2.0's Govern function with five pillars:

- **GV.1 AI Risk Policy & Appetite:** Acceptable use, prohibited use, SL-T targets, approval gates
- **GV.2 Roles & RACI:** OT Owner, AI Product Owner, MLOps Lead, Cybersecurity Lead, Process Safety, Data Steward, Compliance
- **GV.3 AI Supply Chain:** Model/data sources, SBOM/MBOM, vulnerability posture
- **GV.4 Oversight & Assurance:** Review board, red-team cadence, certifications
- **GV.5 Records & Auditability:** Evidence repositories, logs retention, traceability

Integrations: ERM, Change/Release Management, Incident Response, Vendor Management, Training & Competence

9. Lifecycle & Workstreams

Plan → Design → Build → Validate → Operate → Monitor → Improve

Six core workstreams span all lifecycle phases:

- **WS-A: Data Governance** – Catalogue, lineage, quality SLOs, versioning, access controls
- **WS-B: Model Engineering** – Reproducible training, feature store, model registry, test suites
- **WS-C: Secure MLOps** – Secrets, CI/CD, policy-as-code, signed artefacts, deployment control
- **WS-D: Operations** – Telemetry, drift detection, health SLOs, rollback
- **WS-E: Safety & Cyber** – Hazard analysis, zones/conduits, SCAI mapping, ICS hardening
- **WS-F: Assurance** – XAI packs, red teaming, audits, certificates

10. Controls Catalogue (IEC 62443 mapped, AI-specific)

IEC 62443 Security Requirements (SRs):

- **SR 1** – Identification & Authentication (API authN, signer identity)
- **SR 2** – Use Control (RBAC, change control)
- **SR 3** – System Integrity (signed datasets/models, checksums, poisoning detection)
- **SR 4** – Data Confidentiality (TLS, logging, drift detectors)
- **SR 5** – Restricted Data Flow (zones/conduits, egress control)
- **SR 6** – Timely Response (playbooks, auto-quarantine)
- **SR 7** – Resource Availability (HA, rate limits, safe degradation)
- **SR 8** – Audit/Accountability (tamper-evident logs, model cards, XAI packs)

AI-Specific Control Families:

- **DG** – Data Governance (lineage, provenance, PII handling)

- **MR** – Model Robustness (adversarial training, perturbation tests)
- **EX** – Explainability (XAI methods, decision rationales)
- **DM** – Drift Monitoring (statistical tests, performance SLOs, retrain triggers)
- **PS** – Process Safety (SCAI linkage, safe-state enforcement, HIL overrides)

11. Assessment Method (6-Step Process)

The I-CAIRA assessment follows six systematic steps:

Step 5: Integrated Asset & AI Component Inventory
Step 6: Consequence Analysis (Cyber-Physical with SCAI)
Step 7: AI-Specific Likelihood & Threat Modelling
Step 8: AI-Refined Risk Scoring
Step 9: Mitigations & SL-T Justification
Step 10: Continuous Assurance

12. Risk Scoring Rubric

$$\mathbf{RISK = L \times C \times (1 + f(R, X, A, B) / k)}$$

Risk bands: Low (1-5) → Moderate (6-12) → High (13-20) → Extreme (>20), mapped to SL-T 1-4

13. Maturity Model (L0–L4)

L0 Ad-hoc: No inventory, manual deployments, no drift monitoring

L1 Basic: Inventory exists, gated releases, basic logging

L2 Managed: Model registry, CI/CD, signed artefacts, drift SLOs

L3 Integrated: Adversarial testing, XAI packs, automated rollback, SL-T linked controls

L4 Assured: Continuous assurance, independent red teaming, external certification, dynamic risk register

14. Metrics, KRIs & SLOs

- **Model Health:** MAPE/AUC delta, drift D-statistic, data freshness, retrain MTTU
- **Security:** % models with signed lineage, adversarial test pass rate
- **Safety:** # HIL overrides, time in safe-state, nuisance trip rate
- **Governance:** % models with current XAI pack, audit findings closed, SBOM coverage

15. Implementation Roadmap (90-Day Starter)

Days 0–30 (Discover): Stand up inventory, choose toolchain, identify top AI use cases

Days 31–60 (Pilot Controls): Create XAI packs, enable signed artefacts, establish monitoring SLOs

Days 61–90 (Institutionalize): Approvals board live, map SL-T to controls, publish dashboards, schedule red-team

16. Deliverables & Templates

9. AI-OT Asset & Model Register
10. Integrated Risk Scenario Matrix
11. Mitigation Plan with control mapping
12. XAI Pack Template
13. Assurance Checklist
14. Incident Playbooks

17. Roles & RACI

Clearly defined roles: OT System Owner (Accountable for SL-T), AI Product Owner (Accountable for model function), MLOps Lead (Responsible for pipelines), Cybersecurity Lead (Responsible for SR mappings), Process Safety Engineer (Consulted on SCAI), Data Steward (Responsible for provenance), Compliance/Audit (Informed, conducts checks)

18. Policies

- AI in OT Acceptable Use
- Data & Lineage
- Model Release
- Monitoring & Response

19. Annexes

- Glossary
- Reference mappings (IEC 62443 SRs ↔ AI control families)
- Example risk scenarios
- Evidence checklists

20. Comparative Landscape & Alignment

I-CAIRA positions as an OT-specific overlay that fuses PDPC + NIST AI RMF + EU AI Act with IEC 62443/ISA TR84. Include detailed control mapping, overlay guidance, and audit traceability artifacts.

21. V1.1/V2.0 Recommendations

Three Core Areas for Evolution:

1. Enhancing Technical Rigor: Adversarial Resilience & XAI Artifacts

- Mandate Adversarial Retraining and Adaptive Stress Testing methodologies
- Strengthen XAI Audit Pack requirements: algorithm readability, model behaviour, consequence analysis
- Link Bias/Performance Degradation directly to Data Governance with forensic provenance analysis

2. Strategic Governance & Continuous Assurance

- Frame MLOps as Compliance Automation—the technical implementation for achieving IEC 62443 SRs
- Define Dynamic Risk Register with auto-updating Likelihood scores based on MLOps telemetry

3. Commercial Positioning and Pioneer Status

- Position MeruEPC as an 'Anticipator'—pioneering organization ready for autonomous/agentic AI
- Targeted use case integration: AiPSO (upstream optimization), autonomous robotics (field services)
- Reference formal certification pathway (HITRUST AI Security Assessment)

About MeruEPC Pte. Ltd.

MeruEPC
SECURE. SMART. HUMAN

MeruEPC Pte. Ltd. is a Singapore-based cybersecurity and engineering consulting firm pioneering the convergence of Operational Technology security, Artificial Intelligence governance, and industrial safety. Through our 'Secure. Smart. Human.' philosophy, we deliver solutions that enhance both security and operational excellence.

Why MeruEPC?

- ✓ **Pioneer Positioning:** We are anticipators, not reactors—building resilience before threats materialize
- ✓ **Standards Leadership:** Contributing to the development of international AI security frameworks
- ✓ **Industry-Specific Expertise:** Deep understanding of upstream oil & gas, manufacturing, energy, and critical infrastructure
- ✓ **Comprehensive Capability:** From strategy and assessment to implementation and continuous assurance

Contact Information:

Email: vijay@meruepc.com

Website: www.meruepc.com

Head Office: Singapore

Regional Presence: UAE • India • Malaysia • Philippines

Thank You

Join us in pioneering the future of secure industrial AI

For consultation inquiries:

vijay@meruepc.com

https://www.linkedin.com/in/vijay-s-1594386/