Deliverable D3.4

Adversarial attack prevention and mitigation report

30 August 2026

ITEA Project No: 22006

sintra-ai.eu

Abstract:

This report is the direct output of Task 3.4 which aims to protect the AI systems developed in the project from attacks that could compromise their performance and trustworthiness, such as poisoning attacks, evasion attacks, and Trojan attacks.

DOCUMENT VERSIONS

Version no	Date	Changes
1.0	2026-03-01	Initial draft, document outline
v2026.08.30	2026-08-31	Final version

CONTENTS

Introduction

Purpose of the Report

This report, Deliverable D3.4, is the output of Task 3.4 of the SINTRA project. Its primary purpose is to detail strategies for protecting the project's Artificial Intelligence (AI) systems' performance and trustworthiness against adversarial attacks that could compromise their performance, reliability, and overall trustworthiness. The document specifically addresses key threat vectors, including data poisoning, evasion, and Trojan attacks, by outlining comprehensive techniques for prevention (through adversarial training and data augmentation), and mitigation (through detection-based methods and model hardening). This ensures the developed AI systems maintain robust operation in hostile environments.

Overview of Adversarial Threats

Adversarial attacks are intentional manipulations designed to compromise the performance and trustworthiness of AI models. These threats are broadly categorized into three categories based on the point of influence in the AI lifecycle:

Poisoning attacks

These target the training data to degrade model performance or introduce backdoors, often through techniques such as label flipping or clean-label attacks that poison the training data itself.

Evasion attacks

These occur during the model's inference phase, where inputs are subtly disturbed, which often is not detectable by humans, to cause the model to render incorrect predictions.

Trojan or Backdoor attacks

These attacks involve embedding a hidden malicious function during training that is activated only by a specific predefined input condition or trigger. For visual models for example, the model could ignore people holding weapons, only if they're also wearing a neon green sweater and red pants.

This specialized field of study, known as Adversarial Machine Learning (AML), is critical for developing robust defenses against these sophisticated threats.

FOUNDATIONS FOR PREVENTION AND MITIGATION

Insights from Task 3.2

Foundations for Prevention and Mitigation

1. SINTRA D3.2 AI Detection Inventory & Threat Scope

The SINTRA Task 3.2 framework targets a wide spectrum of anomalies and threats across critical infrastructures:

- **Airport CCTV & Radar Analytics:** Real-time fall detection (YOLOv7 Pose + ByteTrack), abandoned luggage classification (YOLOv11s + ByteTrack), fire and smoke recognition (YOLOv7 fine-tuned with Nano Banana synthetic data), and anonymous human tracking (mmWave radar + SA-mPoint DBSCAN + 6D Kalman filtering).
- **Harbour Safety & Surveillance:** Suspicious scene classification (pose-based XGBoost + SHAP explainability), aerial human-action anomaly detection (UAV-recorded video with human-centric tubelet embeddings + modified UR-DMU memory-based loss tested on the custom Sky-VAD dataset for looting, container theft, and physical altercations), and continuous acoustic monitoring (Sorama CNN autoencoders operating on Mel-spectrograms at truck parking and Rode Vaart maritime facilities).
- **Perimeter & Infrastructure Security:** Fence detection with glare compensation (OneFormer semantic segmentation + YOLOv8/10 + HSV/CLAHE preprocessing), multi-view perimeter protection (reconciliation of complementary camera views with Sensoan approach/tampering sensors), and sector access tracking.

2. Actionable Adversarial Mitigation Strategies Derived from D3.2

1. Neutralizing Fleeting "Frame-Injection" and Pixel Perturbation Attacks via Temporal Validation Chains

- **The Vulnerability:** Standard deep learning classifiers are vulnerable to localized pixel-level adversarial perturbations (evasion attacks) or brief frame-injection exploits designed to force misclassifications in isolated video frames (e.g., rendering an intruder "invisible" or triggering false alarms).
- **D3.2 Mitigation Paradigm:** Across all critical SINTRA video analytics modules, raw model outputs are never trusted in isolation. The system embeds a **Common Tracking and Temporal Validation Layer** that utilizes ByteTrack's two-stage association logic to maintain consistent target identity across occlusions, rapid motion, and detection failures.
- **Prevention Strategy:** SINTRA models enforce rigorous temporal validation constraints before generating alerts:

- Fall Detection: Uses a sliding-window biomechanical feature check (torso angle, hip velocity, vertical head-ankle distance) combined with a mandatory hold-time validation period.
- Abandoned Luggage: Employs a state-machine that tracks ownership-association, relative spatial proximity, and velocity/direction consistency over a sustained separation-duration threshold.
- Fire & Smoke: Utilizes an N-of-M frame persistence check and regional motion continuity to separate transient light reflections/glare from sustained combustion.

- **Adversarial Hardening:** Implementing multi-frame sliding-window constraints and track-continuity verification prevents transient, single-frame adversarial pixel manipulations from propagating into system-level alerts, effectively raising the attacker's cost to requiring sustained multi-frame physical exploits.

2. Countering Contextual "Background Bias" and Targeted Hallucinations via Spatial Patch-Level Localisation

- **The Vulnerability:** Weakly-supervised anomaly detectors trained on video-level labels often suffer from "background bias". Instead of learning the anomalous action itself, they learn to correlate anomalies with specific backgrounds, scenery, or visual contexts. An adversary can exploit this by subtly manipulating the background scenery to trigger false alerts (targeted denial-of-service) or mask active intrusions.
- **D3.2 Mitigation Paradigm:** SINTRA integrates the **Sparse Spatio-Temporal Detection and Localisation (SST-WSVADL)** framework. While standard detectors output a single temporal score, SST-WSVADL incorporates a patch-level branch backed by a motion-aware term that penalizes and discards background patches as training progresses, without requiring manual spatial annotations during training.
- **Prevention Strategy:** This dual-branch architecture generates an anomaly score alongside a **localized map of patch scores**. When an alert is flagged, the system visualizes exactly which image region (e.g., the cargo vehicle or the specific fence line being handled) drove the detection.
- **Adversarial Hardening:** Providing spatially grounded anomaly maps enables security operators to immediately identify and audit background-biased false triggers. If the model generates an anomaly alert but the spatial score map highlights an unmanipulated background area rather than the physical subject, the operator can flag the alert as a potential adversarial context exploit or model-drift issue.

3. Defeating Single-Sensor Evasion via Heterogeneous Physical Fusion and Multi-View Spatial Reconciliation

- **The Vulnerability:** Sophisticated intruders can employ physical-world evasion tactics, such as wearing thermal-cloaking apparel, using physical adversarial camouflage patches, or executing acoustic masking (silencing or saturating microphones) to bypass a specific surveillance sensor.
- **D3.2 Mitigation Paradigm:** SINTRA's Port of Kemi perimeter monitoring platform implements a **multi-sensor, multi-view fusion framework**. It combines co-located heterogeneous modalities: two perimeter cameras looking at the same fence line from complementary viewpoints, outdoor audio channels, and Sensoan-provided physical vibration and approach sensors.
- **Prevention Strategy:** Rather than treating sensors independently, the platform executes a **detection-fusion step** that reconciles observations across views into a single canonical event. Detections are spatially aligned and projected onto a common ground plane using known camera geometry as the primary cue, with appearance-based re-identification acting solely as a secondary tiebreaker.
- **Adversarial Hardening:** This architecture restricts the efficacy of single-modality physical attacks. An intruder wearing a visual adversarial patch might evade one camera's edge classifier, but their presence is captured by the second camera's complementary viewpoint, the Sensoan vibration/approach triggers, or acoustic anomaly detection. Reconciling these signals spatially on a shared ground plane prevents bypasses of individual, compromised sensors.

4. Mitigating Insider-Assisted Asset Theft through Dual-Domain (Physical-Digital) Mismatch Detection

- **The Vulnerability:** Insider threats or compromised digital accounts can manipulate inventory databases, terminal gate logs, or access-control registries to digitally authorize illegal asset exfiltration, making the theft appear completely legitimate on paper.
- **D3.2 Mitigation Paradigm:** SINTRA's sector access monitoring and retail shop security modules explicitly combine physical sensory observations with digital database transactions. In Kemi Harbour, the system monitors vehicle and equipment trajectories, dwell-times, and sector transitions. In retail shops, YOLO object detectors and DeepSORT trackers monitor physical product handling at Point-of-Sale (POS) lanes.
- **Prevention Strategy:** The core indicator of a critical anomaly is the **mismatch between physically observed movements and the digital authorization record** within the operational databases. If a truck exits a sector or a product is moved past checkout, but

the operational log contains no corresponding authorization, an alert is triggered immediately.

- **Adversarial Hardening:** This establishes a physical-digital security loop. An adversary cannot bypass security simply by compromising digital databases or cloning credentials; they must also physically spoof the real-time AI tracking systems, which are continuously monitoring trajectories and verifying actions against digital logs.

5. Securing the Edge AI Platform against Firmware Compromise, Prompt Injection, and Sensor Spoofing

- **The Vulnerability:** The AI edge platform itself is an active attack target. Sensors can be physically tampered with, frozen, or compromised at the firmware level; microphones can be masked or saturated; and the natural language reporting models can be targeted with adversarial prompt injections.
- **D3.2 Mitigation Paradigm:** SINTRA's **Shared Harbour Reporting Platform** utilizes locally deployed, quantized open-source LLMs (Llama, Qwen, or Mistral) running in a strictly air-gapped, on-premises environment to comply with security regulations and prevent external data leaks. It builds defense directly into the reporting pipeline:
    - **Sensor Trust & Reputation Scoring:** The platform applies trust- and reputation-scoring at the individual sensor-node level to support zero-trust segmentation of the port-edge network. If a sensor node exhibits symptoms of compromise—such as frame-freezing, calibration drift, silenced microphones, or replayed approach triggers—its reputation score collapses, and its data is isolated or ignored.
    - **LLM Self-Monitoring:** The platform treats the reporting model itself as a monitored asset, continuously tracking model output for "confidence collapse," repeated implausible outputs, or markers of prompt injection.
    - **Traceable Evidence Chains:** To defeat hallucinations and sensory prompt injections (where an attacker crafts visual or acoustic triggers to manipulate the reporting LLM), SINTRA restricts reports to a **fixed schema**. Every statement in the report is bound to an explicit, verifiable evidence chain consisting of direct, timestamped detector outputs and raw evidence clips.
- **Adversarial Hardening:** Local air-gapped execution eliminates network-based exfiltration vectors, while node-level reputation scoring and strict, schema-enforced traceability prevent compromised hardware or manipulated sensory inputs from hijacking the automated SOC reporting pipeline.

Insights from Task 3.3

Assessing the baseline robustness and resilience properties of the current AI systems.

1. Unified Threat Taxonomy and Lifecycle Attack Surface

The SINTRA D3.3 report outlines that multi-modal AI systems deployed in critical infrastructures (such as seaports, transportation hubs, and cyber-physical security systems) must be secured across their entire operational lifecycle. Rather than viewing security through a singular lens, D3.3 classifies the adversarial attack surface into distinct vectors:

- **Inference-Stage Evasion Attacks:** Exploiting the mathematical boundaries of deep neural networks using adversarial examples—carefully calculated, sub-visual perturbations to input data (such as RGB camera frames, audio spectrums, or network traffic logs) that cause models to misclassify threats or completely miss critical anomalies.
- **Training-Stage Data Poisoning:** Malicious injection or modification of training data to introduce backdoors, establish system biases, or systematically blind the AI from recognizing specific threat signatures (e.g., a specific vehicle class or intrusion pattern). This represents a high-impact insider or supply-chain risk that remains dormant until deployment.
- **Model and Data Extraction:** Utilizing repeated queries via public or exposed APIs to reconstruct model weights (model extraction) or reverse-engineer sensitive training data and operational logs (model inversion), facilitating the development of offline evasion strategies.
- **Emerging Agentic and Prompt Exploitation:** Exploiting autonomous, long-running AI agents holding elevated system permissions. Attackers can leverage prompt injection, context manipulation, and unauthorized tool usage to trigger autonomous execution of harmful actions and lateral network movement at machine speed.

2. SINTRA Foundations for Multi-Modal AI Resilience

Traditional AI models degrade rapidly when individual sensory feeds are compromised or unavailable. SINTRA addresses this baseline vulnerability by implementing and analyzing four

state-of-the-art resilience frameworks to guarantee operational continuity under active sensor denial, degradation, or coordination attacks:

- **Modality-Incomplete Adaptation (RADAR Framework):** Standard early- or late-fusion systems fail if a sensor goes offline due to physical tampering, network failure, or environmental masking. SINTRA leverages the RADAR framework, which utilizes a *HyperNetwork* combined with a *hybrid pseudo-module*. This allows the anomaly detector to dynamically adapt to incomplete modality subsets at runtime, minimizing performance loss when specific sensors (such as LiDAR or cameras) are unavailable.
- **Uncertainty-Driven Feature Distillation (AVadCLIP Framework):** In audio-visual surveillance applications, the absence of one channel (e.g., camera occlusion or microphone sabotage) typically blinds joint-modality detectors. SINTRA evaluates AVadCLIP, which uses real-time uncertainty quantification to route information and dynamically synthesize missing representations (e.g., generating missing acoustic indicators from visual-only feeds) to maintain situational awareness.
- **Dynamic Reliability Weighting (RobustA Framework):** Modality streams are rarely either 100% functional or completely offline; they often suffer from varying degrees of environmental noise or active perturbation. SINTRA integrates RobustA to benchmark corrupted streams and apply dynamic reliability weighting at the decision boundary, reducing the influence of corrupted channels while prioritizing stable sensors.
- **Certified Multi-Modal Security (MMCert):** To establish formal bounds on model robustness, SINTRA utilizes the MMCert framework. MMCert provides the first *certified defense* against multi-modal adversarial attacks, mathematically proving that the AI's final classification remains stable under bounded, cross-sensor coordinate perturbations.

3. Sensor-Platform Robustness and Quality Optimization

A fundamental pillar of SINTRA's trustworthy AI framework is ensuring the baseline physical-to-digital integrity of sensory inputs. The project implements specific hardware and preprocessing optimizations to defend downstream classifiers from "garbage-in, garbage-out" failures caused by environmental noise or physical spoofing:

- **Thermal Modality Optimization:** Instead of standard image-quality metrics (like signal-to-noise ratio), SINTRA defines task-relevant *objective detectability metrics*

specifically for thermal-based object classification. It enhances the thermal pipeline using:

  - *Calibration procedures* to stabilize radiometric responses and correct sensor non-uniformities.
  - *Refined gain map estimation* to eliminate fixed-pattern noise and spatial artifacts in low-signal environments.
  - *Targeted noise reduction* to preserve critical edge details while filtering out thermal clutter.

- **RGB Low-Light Preservation:** To complement thermal arrays, RGB sensors employ adaptive exposure control, noise-aware sensor profiling (preventing high-frequency noise amplification), and dynamic range optimization. This preserves color consistency and semantic details necessary for human-in-the-loop validation.
- **mmWave Radar as a Resilient Fallback:** When optical systems are completely blinded (by dense fog, snow, smoke, or physical lens spray) or restricted due to strict GDPR/privacy regulations, millimeter-wave (mmWave) radar serves as an independent, privacy-preserving redundancy channel.
  - SINTRA's radar processing pipeline bypasses heavy deep learning requirements by leveraging *five-dimensional (5D) clustering* (spatial coordinates, velocity vectors, motion characteristics, and intensity).
  - It employs *intensity-based filtering* to isolate human-specific point clusters while suppressing multipath reflections, ghost targets, and environmental clutter caused by massive metallic infrastructure.

4. System-Level Trustworthy AI Governance and Safeguards

The SINTRA framework builds on the European Commission's guidelines for Trustworthy AI (lawful, ethical, and robust) to construct a layered, secure decision-support architecture:

- **Explainable AI (XAI) as an Anti-Fatigue Safeguard:** Black-box neural networks often trigger cryptic alerts, leading to security operator fatigue or enabling slow, low-rate adversarial drift to go unnoticed. By utilizing SHAP and LIME, SINTRA provides visual and logical rationales behind alerts, allowing human operators to quickly audit anomalies, trace feature weights, and confirm whether an alert is a genuine physical threat or a sensor malfunction.

- **Traceability and Forensics:** To maintain legal accountability and meet stringent regulatory requirements in transport infrastructure, SINTRA maintains comprehensive, immutable data-provenance and processing logs, tracking every stage from raw sensor ingestion to the finalized inference output.
- **Human-in-the-Loop Validation:** SINTRA's platform architecture enforces human-on-the-loop verification. AI agents and anomaly models serve strictly as intelligent decision-support utilities, ensuring that final critical responses (such as alarms, gate closures, or security dispatches) are validated by authorized personnel, preventing automation bias or exploit loops.
- **Continuous Validation and Monitoring:** Because operational environments are susceptible to concept drift and evolving adversarial tactics, the framework incorporates regular bias assessments, automated red-teaming, and continuous validation pipelines to retrain models and preserve high baseline detection accuracy throughout the system's operational lifecycle.

Tailoring Strategies

Aligning prevention and mitigation approaches to the specific needs of the project's use cases requires a risk-based tailoring of defense mechanisms. Given that SINTRA deployments span diverse critical infrastructure—from maritime acoustic monitoring to airport CCTV and perimeter physical security—a "one-size-fits-all" defense strategy is insufficient.

Based on the threat taxonomies established in Task 3.3 and the operational detection inventory from Task 3.2, SINTRA applies the following tailored defense strategies:

- **High-Stakes Visual Analytics (Airports/Harbours):** For real-time video modules (e.g., YOLO-based fire/smoke detection, pose-based anomaly detection), defense prioritizes temporal validation chains and spatial patch-level localization. These defenses are tailored to neutralize fleeting, single-frame pixel perturbations by enforcing sliding-window persistence checks and isolating spatial alerts from background noise, effectively countering "background bias" exploits.
- **Multi-Modal Critical Infrastructure (Perimeter/Gateways):** Systems relying on fused sensory data (e.g., radar, optical, and physical sensors) utilize heterogeneous physical fusion. By reconciling observations on a shared ground plane, the platform ensures that an adversary cannot bypass system security through single-modality evasion (e.g., thermal-cloaking or adversarial patches), as physical presence is verified across redundant, complementary sensor views.

- **Log-Integrated Asset Monitoring:** In modules where physical movement interacts with digital registries (e.g., Port of Kemi inventory logs, retail POS transactions), defense is tailored through Dual-Domain (Physical-Digital) Mismatch Detection. This mitigates insider threats by flagging unauthorized exfiltrations where physically observed trajectories lack corresponding digital authorization.
- **Edge-Platform Resilience:** For reporting and analysis nodes (e.g., local air-gapped LLMs), the focus shifts to sensor-node reputation scoring and traceable evidence chains. This protects the autonomous pipeline against firmware compromise and prompt injection by ensuring that any automated report is bound to a verifiable, evidence-based schema, rather than trusting the model output in isolation.

By applying these tailored strategies, SINTRA ensures that defense mechanisms directly address the highest-priority attack surfaces inherent to each operational environment, rather than applying generic hardening to all system components.

Adversarial Training Techniques (Prevention)

Proactive Adversarial Training

Methodologies for exposing models to attacks during the training phase to build robustness.

Adversarial training is a prevention-oriented approach for improving the robustness of AI systems against malicious manipulation. This is especially important because AI models are used for anomaly detection, threat identification, situational awareness, and cyber-physical security monitoring in critical infrastructure environments such as seas ports, airports and transport systems in general.

Deep learning models can be vulnerable to small, carefully designed perturbations that cause incorrect predictions while remaining difficult for humans or operators to notice. Goodfellow, Shlens, and Szegedy introduced the Fast Gradient Sign Method and showed that adversarial examples can transfer across models, making them a practical security concern rather than only a theoretical weakness. NIST also treats adversarial machine learning as a lifecycle risk, covering attacks against training data, model behaviour, inputs, outputs, and deployment environments.

Adversarial training should therefore be understood as a layered prevention method. It does not replace monitoring, access control, secure data pipelines, or system-level anomaly detection. Instead, it strengthens the AI model itself so that the model is less fragile when exposed to manipulated images, sensor readings, logs, communication data, or cyber-physical event streams.

Training with adversarial examples means deliberately exposing the AI model to attack-like inputs during training. Instead of training only on clean data, the model is trained on both normal samples and adversarially modified samples. This teaches the model to make more stable decisions under hostile or abnormal conditions.

The classic formulation of adversarial training is based on a min-max optimization problem: the inner process generates difficult adversarial examples, while the outer process updates the model to classify these examples correctly. Madry et al. formalized this as robust optimization and argued that Projected Gradient Descent adversarial training provides a strong defence against first-order adversaries. This is particularly relevant for critical infrastructure AI systems because attackers may not need to fully compromise the system; they may only need to introduce small manipulations that shift model outputs, suppress alerts, or create false alarms. Adversarial examples can be generated for several data modalities:

- Image and video data. Surveillance models can be attacked through small visual perturbations, lighting manipulation, occlusion, compression artefacts, stickers,

camouflage, or object appearance changes. In a port environment, this could affect detection of unauthorized persons, vehicles, drones, safety violations, or abnormal movement patterns.

- Sensor and telemetry data**.** Environmental sensors, access-control logs, operational telemetry, and IoT streams may be manipulated through value spoofing, delayed messages, missing values, or injected abnormal readings. Adversarial training can simulate such conditions so that models become less sensitive to unrealistic but strategically chosen changes.

- Network and communication data. AI models used for cyber anomaly detection may be exposed to adversarially modified traffic patterns, packet timing changes, protocol-level distortions, or log manipulation. This improves resilience against evasion attempts where attackers try to hide malicious behaviour inside normal-looking traffic.

A practical adversarial training pipeline should include threat modelling, adversarial example generation, mixed clean and adversarial training, robustness evaluation, retraining, and continuous monitoring. The adversarial samples should not be random only. They should reflect realistic attacks against the operational environment. MITRE ATLAS is useful here because it provides a knowledge base of adversary tactics and techniques against AI-enabled systems based on real-world observations.

However, adversarial training must be evaluated carefully. Carlini and Wagner demonstrated that some previously proposed defences were much weaker than expected when tested with stronger attacks. Athalye, Carlini, and Wagner later showed that some defences create “obfuscated gradients,” which can give a false sense of security because they appear robust only against weak or poorly adapted attacks. Therefore, SINTRA models should be tested using multiple attack types, including white-box, black-box, transfer-based, and adaptive attacks.

Input Data Augmentation

Input data augmentation is a prevention technique that increases the diversity of training data by applying controlled transformations to existing samples. The goal is to prevent the model from becoming too dependent on narrow patterns that attackers can exploit. Augmentation improves generalization and makes it more difficult for attackers to craft inputs that exploit fragile decision boundaries.

For image and video analytics, augmentation may include rotation, scaling, cropping, brightness changes, contrast variation, blur, occlusion, compression, rain, fog, snow, darkness, camera vibration, and viewpoint changes. Shorten and Khoshgoftaar's survey identifies geometric transformations, colour-space augmentation, kernel filters, random erasing, adversarial training, GAN-based augmentation, and other techniques as major data augmentation methods in deep learning.

For maritime and port-related use cases, this is important because real operational environments are rarely clean. Cameras may be affected by arctic weather, poor lighting, moving vehicles, reflective surfaces, snow, rain, fog, darkness, or temporary obstruction. If models are trained only on ideal images, they may fail under normal operational stress or become easier to attack.

For time-series and sensor data, augmentation may include jittering, scaling, time warping, masking, random dropout, missing-value simulation, and synthetic fault injection. These methods can simulate sensor degradation, communication delay, temporary outages, or adversarial manipulation. This is relevant for multi-modal AI systems where camera feeds, environmental sensors, access-control data, and operational logs must be interpreted together.

Noise and Perturbation Injection

Noise and perturbation injection strengthens AI models by exposing them to controlled uncertainty during training. The purpose is to make the model less sensitive to small input changes and more stable under both natural disturbances and adversarial manipulation.

At the input level, noise can be added to images, sensor values, network traffic features, and time-series data. Examples include, missing values, timestamp perturbation, signal attenuation, and random masking. For sea port environments, this can simulate camera noise, poor weather, sensor drift, communication interference, or incomplete data streams.

Detection-Based Methods (Mitigation)

Identifying Adversarial Examples

Detecting adversarial examples involves analyzing inputs or internal model states for patterns that deviate significantly from the learned distribution of benign data. Several approaches are employed in the SINTRA framework:

- **Statistical Distribution Testing:** Comparing the statistical properties of incoming data against a baseline of clean training data. Significant deviations in feature activations or input entropy can signal the presence of malicious perturbations.
- **Uncertainty Estimation:** Leveraging techniques like Monte Carlo Dropout or Deep Ensembles to measure the model's confidence in its predictions. Adversarial examples often push inputs into regions of high uncertainty where the model's decision boundaries are brittle, allowing the system to flag suspicious predictions rather than acting on them.
- **Dedicated Detector Models:** Utilizing secondary, lightweight models specifically trained to distinguish between clean samples and adversarial inputs. By treating adversarial detection as a binary classification problem, the system adds a defensive layer that can be updated independently of the primary task-specific model.

System-Level Anomaly Detection

Monitoring the AI systems themselves to detect abnormal or suspicious behavior indicative of an ongoing attack.

System-level anomaly detection goes beyond input validation by monitoring the operational integrity of the AI system as a whole. This focuses on identifying behavioral patterns that are inconsistent with legitimate usage:

- **Operational Drift Monitoring:** Continuously tracking model output statistics, latency, and resource consumption. Sudden shifts, such as an unexplained spike in inference time or a statistically improbable surge in specific class predictions, may indicate an evasion or poisoning attack.
- **Log-Based Behavioral Analysis:** Correlating AI system outputs with external system logs (e.g., access control, network traffic, sensor heartbeat). An anomaly is flagged if the AI's predictions are physically or logically inconsistent with the surrounding environment, such as a high-confidence security alert generated without corresponding sensor triggering.

- **Cross-Validation and Redundancy Checks:** Comparing predictions from independent, diverse model architectures or modalities. If one model's output drastically diverges from the consensus or established ground truth (e.g., a camera detecting an intruder while radar and vibration sensors report silence), it triggers an investigation into potential model compromise or sensor spoofing.

Model Hardening and Resistance

Hardening Strategies

Structural techniques designed to improve the model's innate resistance to adversarial exploits. Key approaches include defensive distillation, which reduces gradient sensitivity by training models to predict probability distributions rather than hard labels; robust architecture design, incorporating activation functions that are less prone to gradient vanishing or exploding under perturbation; and adversarial pruning, which removes neurons or pathways that are highly sensitive to adversarial triggers, thereby isolating and neutralizing potential backdoors or Trojan pathways.

Active Mitigation

Response protocols initiated immediately upon the detection of adversarial activity. Strategies include dynamic threshold adjustment, where the system automatically increases confidence requirements for alerts if suspicious data patterns are flagged; fail-safe model switching, which diverts traffic to a secondary, high-reliability model—often a simpler, non-learning-based heuristic—if the primary model exhibits signs of compromise; and integrated incident response, where detected anomalies trigger alerts to Security Operations Centers (SOC) for human-in-the-loop verification and manual override, ensuring that automated systems never act on high-risk, unverified data.

Conclusion and Future Recommendations

It is clear that a robust, multi-layered defensive framework is necessary for AI-enabled critical infrastructure. By integrating adversarial training, detection-based mitigation, and model hardening, the project demonstrates that securing AI in high-stakes environments—such as maritime ports, airports, and physical perimeter monitoring—requires moving beyond standalone model optimization. A "Defense-in-Depth" strategy should be prioritized, where security is embedded into the entire AI lifecycle, from raw sensor ingestion to autonomous reporting and human-in-the-loop decision-making.

Vulnerabilities are not limited to single-point model failures but span a broad surface area, including training data poisoning, inference-stage evasion, and emerging agentic threats. By implementing temporal validation chains, heterogeneous physical fusion, and dual-domain mismatch detection, SINTRA can maintain system resilience even when individual components are compromised or degraded. However, as the adversarial landscape evolves—with threats becoming increasingly sophisticated through adaptive prompt injection and agentic exploitation—static defenses must transition toward dynamic, self-evolving protocols.

To maintain this security posture, we recommend the following strategic priorities:

- **Transition to Automated Red Teaming:** While manual validation is critical, the scale of deployment necessitates the adoption of automated red-teaming pipelines. These should continuously simulate emerging attack vectors (white-box, black-box, and adaptive attacks) against deployed models to identify drift and fragility before they can be exploited.
- **Institutionalize Human-Centric Governance:** The most effective safeguard against adversarial manipulation remains the informed human operator. We recommend further refining Explainable AI (XAI) tools to reduce "operator fatigue." Future efforts should focus on creating intuitive, risk-calibrated dashboards that present not just alerts, but actionable forensic evidence, allowing operators to rapidly distinguish between genuine security threats and adversarial noise.
- **Establish a Continuous Validation Loop:** Model performance is not a static metric; it is subject to concept drift and environmental changes. We recommend establishing a rigorous, schedule-based retraining pipeline that incorporates real-world adversarial telemetry into the training sets. This ensures that the defense mechanisms adapt to new camouflage, spoofing, and evasion techniques observed in the field.
- **Promote Open-Source Security Standards:** As the SINTRA project matures, the methodologies developed herein—specifically regarding multi-modal resilience (RADAR/AVadCLIP) and certified security (MMCert)—should be codified into reusable security modules. Sharing these frameworks will not only enhance the reliability of the

project's own infrastructure but contribute to the broader ecosystem of Trustworthy AI in critical infrastructure.

Ultimately, securing AI should be treated not as a destination, but as a continuous operational capability. By prioritizing the integration of physical and digital security, maintaining rigorous traceability, and empowering human oversight, SINTRA provides a blueprint for the next generation of resilient, trustworthy, and mission-critical AI systems.

ACRONYMS

Acronym	Expanded
AI	Artificial Intelligence
AML	Adversarial Machine Learning
AVadCLIP	Audio-Visual Adversarial Contrastive Language-Image Pre-training
CCTV	Closed-Circuit Television
CNN	Convolutional Neural Network
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
GDPR	General Data Protection Regulation
LIME	Local Interpretable Model-agnostic Explanations
LLM	Large Language Model
mmWave	Millimeter Wave
MMCert	Multi-Modal Certified Security
POS	Point-of-Sale
RADAR	Radio Detection and Ranging
SHAP	SHapley Additive exPlanations
SOC	Security Operations Center
UAV	Unmanned Aerial Vehicle
XAI	Explainable Artificial Intelligence
YOLO	You Only Look Once