Robust and resilience of AI systems report

Multi-modal trustworthy AI analysis anomalies, threats, crime

30 Aug 2026

ITEA Project No: 22006

sintra-ai.eu

ITEA 4

DOCUMENT VERSIONS

Version no	Date	Changes
0.1	10 March 2026	Template created
0.2	14 April 2026	Extended draft introduction
0.3	18 May 2026	Initial contributions from partners
0.9	18 June 2026	Updated version ready for review
0.95	23 June 2026	Updated version for final review
1.0	27 August 2026	Final version ready

CONTENT

1 ACRONYMS

Acronym	Expanded
AI	Artificial Intelligence
AI RMF	AI Risk Management Framework
API	Application Programming Interface
APT	Advanced Persistent Threat
CNN	Convolutional Neural Networks
CPS	Cyber-Physical Systems
DL	Deep Learning
GNSS	Global Navigation Satellite System
GOAD	Game of Active Directory
GPS	Global Positioning System
GUI	Graphical User Interfaces
ICS	Industrial Control Systems
IDS	Intrusion Detection Systems
LIME	Local Interpretable Model-Agnostic Explanations
OT	Operational Technology
ML	Machine Learning
MITM	Man-In-The-Middle
mmWave	Millimeter-Wave
NIST	National Institute of Standards and Technology
RGB	Red Green Blue
RNN	Recurrent Neural Networks
SHAP	SHapley Additive exPlanations
VAE	Variational Autoencoder
XAI	Explainable AI

2 ABSTRACT

Artificial Intelligence (AI) is becoming a fundamental technology for protecting critical infrastructures against increasingly complex cyber, physical, and hybrid threats. Modern security environments generate large volumes of heterogeneous data originating from surveillance systems, industrial sensors, communication networks, radar systems, thermal cameras, operational technology (OT) systems, and external intelligence sources. Multi-modal AI enables the fusion and analysis of these diverse information streams to enhance situational awareness, anomaly detection, threat identification, and crime prevention.

However, the growing dependence on AI introduces new vulnerabilities. Adversarial attacks, data poisoning, sensor manipulation, prompt injection, AI-agent exploitation, and communication disruptions can significantly degrade system performance and compromise operational decision-making. As a result, future AI systems must not only achieve high detection accuracy but also demonstrate robustness, resilience, explainability, and trustworthiness under adverse operating conditions.

This report investigates the state of the art in multi-modal trustworthy AI for anomaly, threat, and crime detection. The report analyzes emerging attack vectors against AI systems, evaluates methods for improving AI robustness and fault tolerance, and examines resilience mechanisms required for critical infrastructure deployment. Special attention is given to cyber-physical systems, critical infrastructure protection, seaport security, and multi-sensor situational awareness environments.

Based on the findings, the report proposes the SINTRA Trustworthy AI Framework, a layered architecture integrating trusted data acquisition, secure data governance, robust multi-modal AI processing, resilient decision support, human oversight, and continuous validation mechanisms. The framework provides guidance for developing AI systems capable of maintaining operational effectiveness despite adversarial attacks, sensor failures, communication disruptions, and degraded environmental conditions.

3 INTRODUCTION

The increasing digitalization of critical infrastructures, transportation systems, and seaport environments has significantly expanded the volume and diversity of data available for security monitoring and operational decision-making. At the same time, cyber threats, physical security incidents, criminal activities, and hybrid attacks have become increasingly sophisticated, making traditional rule-based monitoring approaches insufficient for timely detection and response. Artificial Intelligence (AI) has emerged as a key enabling technology for addressing these challenges by providing advanced capabilities for anomaly detection, threat identification, risk assessment, and decision support.

The SINTRA project aims to develop a trustworthy, interoperable, and privacy-preserving AI platform capable of integrating and analyzing data from multiple heterogeneous sources. By combining information from visual sensors, thermal cameras, radar systems, IoT devices, operational technology (OT) environments, cybersecurity monitoring systems, and other data streams, SINTRA enables a comprehensive understanding of complex operational environments. Such multi-modal analysis improves situational awareness and supports the early detection of anomalies, emerging threats, suspicious behaviors, and criminal activities.

However, the adoption of AI in security-critical environments introduces new challenges. AI systems themselves can become targets of cyberattacks, manipulation attempts, data poisoning, adversarial inputs, and privacy-related concerns. Furthermore, security operators and infrastructure owners must be able to trust AI-generated recommendations and understand the limitations, risks, and uncertainties associated with automated decision-making. Therefore, the development of trustworthy AI requires careful consideration of security, robustness, transparency, explainability, privacy protection, and ethical compliance throughout the system lifecycle.

This deliverable examines the role of multi-modal trustworthy AI in the analysis of anomalies, threats, and criminal activities. It reviews relevant AI technologies, discusses security and trustworthiness requirements, analyses the threat landscape affecting AI-enabled security systems, and evaluates potential attack vectors targeting AI models, data sources, and operational environments. Particular attention is given to the resilience of AI-based monitoring solutions operating in cyber-physical environments where failures or compromises of individual sensors must not jeopardize overall system effectiveness.

The report further explores methods for risk assessment, threat mitigation, and operational resilience. Special emphasis is placed on the use of complementary sensing technologies, such as mmWave radar, which can provide privacy-preserving and robust situational awareness even when optical sensing systems are unavailable, degraded, or restricted by regulatory requirements. Through these analyses, the report contributes to the development of secure, resilient, and trustworthy AI solutions capable of supporting critical infrastructure protection and advanced security operations in complex real-world environments.

3.1 Background

Cyber threats targeting critical infrastructures have become more sophisticated and increasingly combine cyber, physical, informational, and operational elements. Nation-state actors, organized cybercriminal groups, insiders, and hybrid threat actors exploit vulnerabilities across multiple domains simultaneously. Traditional security systems often struggle to detect these complex attack patterns because they rely on isolated data sources and predefined detection rules.

Artificial Intelligence offers significant potential for improving situational awareness and threat detection. By learning complex patterns from large datasets, AI systems can identify anomalous behaviour, predict emerging threats, and support decision-making processes. Multi-modal AI extends these capabilities by integrating heterogeneous data sources into a unified analytical framework.

While AI offers substantial benefits, it simultaneously introduces new risks. Adversarial machine learning, data poisoning, sensor spoofing, prompt injection, model extraction, and AI-agent manipulation have emerged as significant security concerns. Consequently, trustworthiness, robustness, and resilience have become fundamental requirements for AI systems deployed in critical environments.

4 STATE OF THE ART

4.1 AI for Anomaly and Threat Detection

Artificial Intelligence (AI) has become a cornerstone technology in anomaly and threat detection across cybersecurity, industrial control systems (ICS), and cyber-physical environments. Traditional rule-based systems are increasingly replaced or augmented by machine learning (ML) and deep learning (DL) techniques that can model complex patterns and detect previously unseen threats.

Supervised learning approaches, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), are widely used for classification tasks in intrusion detection systems (IDS) and malware detection. However, their reliance on labeled datasets limits their effectiveness in detecting zero-day attacks. To address this limitation, unsupervised and semi-supervised techniques—such as autoencoders, variational autoencoders (VAEs), and clustering methods—are employed to identify anomalies as deviations from learned normal behavior [1].

More recently, graph-based learning and graph neural networks (GNNs) have gained attention for modeling relationships in network traffic and system interactions [2]. These approaches are particularly effective in detecting coordinated or distributed attacks. In cyber-physical systems (CPS), anomaly detection extends beyond digital signals to include physical sensor data. Hybrid models combining physics-based and data-driven approaches have been proposed to improve detection accuracy and interpretability [3].

4.2 Multi-Modal Data Analysis

Modern threat detection systems increasingly rely on multi-modal data, integrating heterogeneous sources such as network logs, sensor measurements, video feeds, and textual reports. Multi-modal learning enables richer representations and improved robustness compared to single-modality approaches. Fusion techniques are typically categorized into:

- Early fusion, where raw data from different modalities are combined,
- Late fusion, where independent models are combined at the decision level,
- Hybrid fusion, combining both approaches [4].

Deep learning architectures such as multimodal transformers and attention-based models have shown strong performance in capturing cross-modal dependencies [5]. These models are particularly useful in security applications where anomalies may manifest differently across modalities. Challenges in multi-modal analysis include data alignment and synchronization, missing or noisy modalities, computational complexity. In robust multi-modal fusion, models can adapt to missing or corrupted inputs while maintaining performance [6].

Recent literature shows a clear shift from unimodal anomaly detection towards multimodal systems that fuse video, audio, text, motion, RGB–3D geometry, or logs plus metrics, but the literature that jointly addresses multimodality, anomaly/threat detection, and security-centred trustworthiness is still relatively sparse. The strongest recent thread does not yet centre on fairness or provenance; instead, it operationalises trustworthiness as robustness to missing/corrupted modalities, resistance to adversarial perturbations, calibrated uncertainty for degraded sensing, and deployment resilience under data or sensor faults. The most relevant, recent works therefore

cluster around four ideas: CLIP/VLM-based multimodal detectors for surveillance [8], efficient RGB–3D industrial anomaly models with explicit deployment constraints [9], modality-incomplete or corruption-robust detectors [10], and multimodal robustness frameworks/benchmarks such as MMCert, MMAD, AnoVox, and updated VAD metrics [13] . The main research gap is that true cyber-resilience evaluation—coordinated cross-modal attacks, prompt injection, sensor spoofing, fail-safe routing, and graceful degradation - lags behind progress in multimodal feature learning [7]. Multi-modal data fusion is also highly valuable during the development and training phases to reduce operational computational complexity. High-resolution LiDAR strictly as a "ground truth" reference to validate mmWave radar models can be highlighted here. This demonstrates how multi-modal data collection can yield a lightweight, single-modality (radar) operational system that is highly robust.

For this review, trustworthiness is taken in the sense of resilience to cyber-security or operational attacks, including adversarial perturbations, sensor failure, missing/corrupted modalities, and robust deployment under imperfect data. That framing aligns best with recent multimodal AD literature: RADAR explicitly studies modality-incomplete industrial anomaly detection, AVadCLIP uses uncertainty-driven distillation for missing audio at inference, RobustA studies corrupted modalities, and MMCert gives the first certified defence against multimodal adversarial attacks [12], . A broader multimodal literature confirms that missing modalities arise from sensor malfunctions, privacy constraints, interference, and transmission failures, i.e. from exactly the kinds of deployment faults that a resilient detector must survive [7].

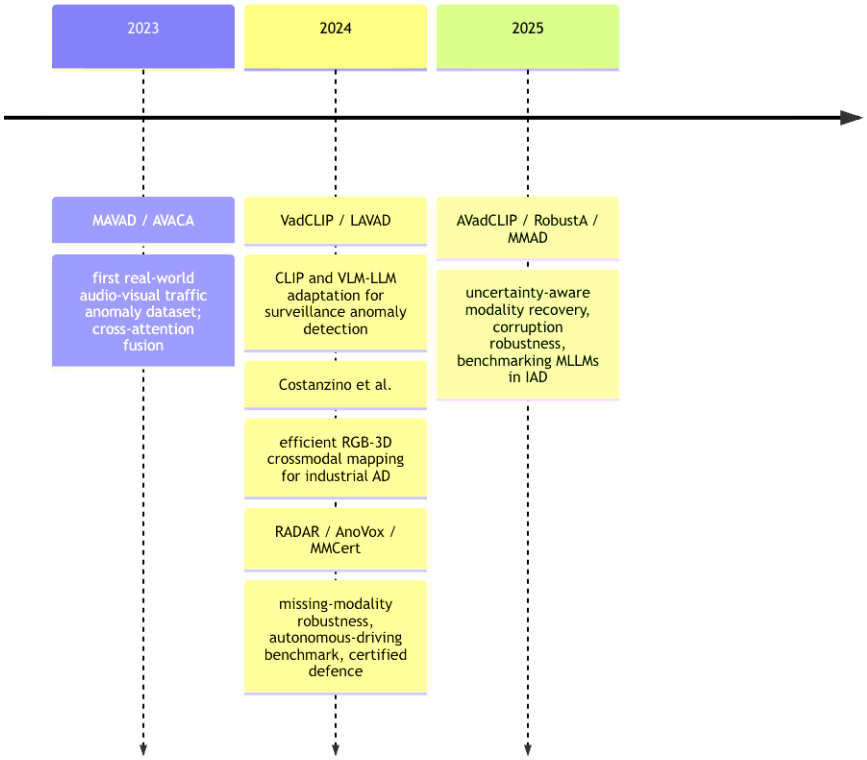

Figure 1 Key recent advances

The timeline in figure 1 shows a progression from “better fusion” towards “graceful degradation and secure deployment”, although robust multimodal AD is still behind mainstream multimodal robustness research [18].

The dominant architecture families are now fairly clear. In surveillance VAD, the centre of gravity has moved to foundation-model adaptation. VadCLIP keeps CLIP frozen and adds a dual branch for coarse visual detection plus fine language-image alignment, reaching 84.51% AP on XD-Violence and 88.02% AUC on UCF-Crime [7]. LAVAD goes one step further and makes VAD training-free, combining off-the-shelf captioning, VLM similarity, and LLM reasoning; on UCF-Crime it reaches 80.28 AUC, and on XD-Violence it outperforms the best unsupervised baseline by 17.03 AUC points [8]. These methods are important because they replace costly retraining with prompt or inference-time adaptation, which is attractive when labelled threat data are scarce or sensitive [7].

Industrial multimodal AD is currently the clearest arena for fault-tolerant architecture. Costanzino et al. [9] use explicit cross-modal feature mapping between RGB and point clouds: anomalies are the inconsistency between what one modality predicts and what the other actually shows. That design is both accurate and deployable: on MVTec 3D-AD, the full model reports I-AUROC 0.954, P-AUROC 0.993, AUPRO@30% 0.971, AUPRO@1% 0.455, while running at 21.755 fps; the tiny variant still gives strong segmentation while reaching 42.818 fps. This is a notable SOTA because security-critical deployment is limited not only by accuracy, but also by memory footprint and latency [9].

The next step is explicit degraded-mode resilience. RADAR reframes MIAD as modality-incomplete detection: rather than assuming all 2D and 3D inputs are present, it uses modality-incomplete instructions plus a HyperNetwork and a hybrid pseudo module so the detector can adapt to missing combinations at train and test time [10]. AVadCLIP does a similar thing in audio-visual VAD with uncertainty-driven feature distillation, learning to synthesise missing audio-visual representations from visual-only inputs when audio is absent or unreliable [17]. RobustA pushes this further by explicitly benchmarking corrupted modalities and using dynamic reliability weighting at inference [12]. Collectively, these papers suggest that the most promising resilient architecture is not a monolithic early-fusion block, but a confidence-aware, modality-adaptive pipeline with graceful degradation [10].

4.3 Security-centred trustworthiness

The most mature security-centred trust axis in this literature is fault tolerance rather than fully fledged adversarial security. Many anomaly papers now test whether the detector survives missing modalities, corrupted streams, or hard-normal shifts; far fewer evaluate coordinated RGB+audio, RGB+LiDAR, or prompt-level attacks. RADAR explicitly shows that standard multimodal industrial methods degrade sharply on missing-modality settings, and AVadCLIP and RobustA make the same point for audio-visual surveillance [12].

The literature on adversarial resilience is therefore still somewhat imported from broader multimodal safety research. MMCert is especially relevant because it is the first certified defense for multimodal models, proving robustness under bounded perturbations across all modalities and outperforming unimodal-certified baselines extended to multimodal settings [19]. For anomaly and threat detection, this is highly consequential: robust multimodal AD should not only score well on clean inputs, but also maintain a provable decision under constrained cross-modal attacks. At present, very few anomaly papers provide such guarantees [13].

Uncertainty quantification is beginning to appear as a practical deployment tool. AVadCLIP does not use uncertainty merely for calibration; it uses it to route information under modality deficiency, which is closer to a fault-tolerant runtime policy than to classical post-hoc confidence estimation [17]. This is precisely the kind of trust mechanism that resilient deployment needs: detect when one sensor is unreliable, reduce its influence, and abstain or fall back when uncertainty remains high. The broader missing-modality literature also supports recovery, retrieval, generation, and ensemble selection as standard design patterns for such degraded-mode inference [17].

Under a broader trustworthy-AI lens, recent multimodal work does discuss explainability, fairness, and provenance, but these are not the main axis of the joint anomaly/threat literature. AnomalyRuler and Holmes-VAD improve reasoning and explanation; multimodal fairness is still comparatively under-studied overall; provenance audits show that multimodal datasets depend heavily on web-crawled and social-media sources with potentially complex restrictions [20]. For your security-focused framing, these issues are secondary to attack resilience, but they still matter for safe deployment in regulated settings. [17]

4.4 Benchmarks and metrics

For surveillance threat detection, the default metrics are still frame-level ROC-AUC and sometimes AP, usually on UCF-Crime and XD-Violence. That remains useful, but recent work argues these metrics are insufficient because they are sensitive to single-annotation bias, do not reward early detection, and fail to expose scene overfitting. Liu et al. [16] therefore propose Prob-AUC/AP, Latency-aware AP, and hard-normal benchmarks such as UCF-HN and MSAD-HN. For multimodal explanation or causation, CUVA adds “what/why/how” annotations and introduces MMEval, which correlates better with human preference for explanation quality than standard text-generation scores [16], [23].

For industrial AD, the important metrics are image-level AUROC, pixel-level AUROC, and AUPRO, especially the stricter AUPRO@1%; recent work also reports fps and memory, which should be treated as first-class deployment metrics rather than mere engineering details [10]. For multimodal MLLM-based IAD, MMAD is currently the most informative benchmark: it spans 39,672 questions, 8,366 images, and seven subtasks, and shows that even strong commercial models only reach about 74.9% average accuracy, still below industrial expectations [14].

Under a resilience-focused definition of trustworthiness, future evaluations should report not only clean AUROC, but also: robust AUROC/AP under attack or corruption, performance drop under missing-modality masks, certified robustness radius/budget, abstention/calibration error under uncertainty, and latency/memory under degraded mode. Current benchmarks only partially support this [10].

5 Trustworthy AI Principles

Trustworthy Artificial Intelligence (AI) is a fundamental requirement for security-critical applications operating in critical infrastructure environments. In domains such as transportation, public safety, and maritime operations, AI systems increasingly support operational decision-making, anomaly detection, threat intelligence, and incident response. Consequently, stakeholders must be able to trust that AI-generated outputs are reliable, secure, explainable, and resilient under both normal and adverse operating conditions.

The European Commission's Ethics Guidelines for Trustworthy AI define trustworthy AI as AI that is lawful, ethical, and robust throughout its lifecycle [24]. While legal compliance and ethical considerations remain important, operational robustness is particularly critical in cybersecurity and critical infrastructure contexts, where incorrect decisions may result in safety incidents, operational disruptions, financial losses, or national security consequences. Trustworthy AI in anomaly and threat detection systems should be built upon following key principles.

Resilience and Robustness

AI systems must maintain acceptable levels of performance despite adversarial attacks, sensor failures, communication disruptions, data corruption, or unexpected environmental conditions. Robustness requires resistance to adversarial manipulation, while resilience focuses on the system's ability to continue operating and recover from failures. In multi-modal environments, resilience is often achieved through sensor redundancy, adaptive data fusion, graceful degradation mechanisms, and continuous monitoring.

Explainability and Interpretability

Security analysts and operational personnel must understand why an AI system has generated a specific alert, prediction, or recommendation. Explainable AI (XAI) techniques improve transparency by revealing the factors and data features influencing model decisions. This is particularly important in threat detection applications, where operators must rapidly assess the credibility of alerts and determine appropriate response actions. Techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) have become widely adopted methods for improving model interpretability and supporting human oversight.

Traceability and Auditability

Trustworthy AI systems should provide complete traceability of data sources, model versions, processing steps, and decision outputs. Traceability enables forensic investigations, regulatory compliance, incident analysis, and continuous improvement of AI models. Auditability ensures that system behavior can be reviewed and validated by internal security teams, regulators, and other stakeholders. Maintaining detailed logs and provenance information is particularly important in critical infrastructure environments where accountability requirements are stringent.

Security and Data Protection

AI systems often process sensitive operational, commercial, or personal data. Therefore, trustworthy AI requires strong cybersecurity controls throughout the data lifecycle, including secure data collection, encryption, access control, authentication, integrity verification, and secure communications. Data governance mechanisms should ensure that information is protected against unauthorized access, manipulation, and disclosure while maintaining compliance with relevant privacy and cybersecurity regulations.

Fairness and Non-Discrimination

Although fairness is most commonly associated with human-centric AI applications, it is also relevant in security systems. AI models should avoid systematic biases that could lead to inaccurate threat assessments, disproportionate targeting of specific groups, or unequal treatment of users and stakeholders. Regular bias assessments and validation procedures should be incorporated into the AI lifecycle to ensure equitable and consistent system behavior.

Human Oversight and Accountability

AI should support, rather than replace, human decision-makers in critical security operations. Human operators must retain the ability to understand, validate, override, and challenge AI-generated recommendations when necessary. Human-in-the-loop and human-on-the-loop approaches help prevent automation bias and ensure that accountability remains with qualified personnel rather than autonomous systems.

Continuous Validation and Monitoring

Trustworthiness cannot be guaranteed through initial testing alone. AI models operating in dynamic environments are susceptible to concept drift, evolving threat landscapes, and changing operational conditions. Continuous monitoring, periodic retraining, performance validation, and

security testing are therefore essential for maintaining trustworthiness throughout the operational lifecycle.

Trustworthy AI Frameworks and Standards

Several international frameworks and standards provide guidance for implementing trustworthy AI systems. These include the European Commission's Ethics Guidelines for Trustworthy AI, ISO/IEC 24028 (Artificial Intelligence – Trustworthiness), the NIST AI Risk Management Framework (AI RMF), and emerging AI governance regulations. Together, these frameworks promote systematic approaches to risk management, transparency, accountability, resilience, and security.

In the context of the SINTRA project, trustworthy AI extends beyond traditional ethical considerations to include cyber resilience, fault tolerance, secure multi-modal data fusion, and operational continuity. The ultimate objective is to ensure that AI-enabled anomaly and threat detection systems remain reliable, explainable, and effective even when operating in adversarial and degraded environments. Such capabilities are essential for supporting critical infrastructure protection and maintaining confidence in AI-assisted decision-making.

5.1 Threats Against AI Systems

By 2026, AI agents have moved from the lab to the real world and completely changed the cybersecurity landscape. These systems can think, use tools, and achieve goals on their own. However, this widespread use creates a major new attack surface. Because these autonomous agents run continuously and often hold elevated permissions, they cross trust boundaries that traditional security systems simply were not built to handle [25]. As a result, these nonhuman workers are vulnerable to new types of attacks like prompt manipulation and automated lateral movement. This dynamic gives attackers the ability to build automated toolkits and strike at machine speed. These fast attacks easily bypass older security controls designed around human behavior [24].

Despite these risks, agentic AI is also becoming a core part of modern defense strategies. To keep up with the volume and speed of modern threats, security teams are deploying their own defensive AI agents. These defenders automate complex tasks like triaging alerts, reverse engineering malware, and guiding incident response[26]. Securing this new environment means shifting from traditional access controls to continuous monitoring of what the AI is actually doing. This involves using Zero Trust principles and dynamic policies to give AI systems strict and temporary permissions [26]. Additionally, security teams are actively using automated red teaming and safe testing environments. They rigorously test and harden these AI agents before deployment to ensure automated defenses can safely stay one step ahead of the attackers [25].

6 THREAT MODEL AND ATTACK SCENARIOS

AI-based security systems operate in adversarial environments where attackers continuously evolve their techniques. The threat landscape includes cyber attackers targeting model vulnerabilities, insider threats manipulating training data, nation-state actors targeting critical infrastructure. The integration of AI into CPS expands the attack surface, as both digital and physical components can be exploited.

6.1 AI-Agent Attack Capabilities

AI agents have shown some serious advantages when it comes to systematic enumeration and massive parallel exploitation. In the December 2025 Stanford ARTEMIS study, a multi agent framework working on an 8,000 host enterprise network took second place overall. It found 9 valid vulnerabilities and outperformed 90% of human professionals at a super efficient cost of $18 per hour [27].

However, there is still a massive gap between the lab and the real world. While agents do great in sanitized environments, they only succeed on about 13% of real world CVE exploitation benchmarks [29]. Agents consistently struggle with multi step Graphical User Interfaces (GUIs) [28]. And even though they are getting better at semantic reasoning for business logic, they currently sit at near a 0% success rate on blind injection vulnerabilities that require timing inference [30]. However, Anthropic's Claude Mythos preview model, which is not publicly released, showed where the state-of-the-art with frontier models are. It autonomously found multiple zero-day vulnerabilities (including a 16 year old flaw in FFmpeg) and chained Linux kernel exploits. This prompted a restricted defensive rollout called Project Glasswing [31].

The market is split right now based on deployment needs and data sovereignty. Commercial Platforms like XBOW and NodeZero focus on continuous execution in production environments. NodeZero, for example, autonomously solved the complex Game of Active Directory (GOAD) benchmark in just 14 minutes. That means it executed a full kill chain 50 times faster than human baselines [32]. XBOW's autonomous agent successfully reached the #1 rank on HackerOne's global leaderboard [33].

Open Source Frameworks like xOffense, MAPTA, and D CIPHER are optimized for research and environments where data sovereignty is the top priority. These frameworks often match commercial effectiveness by using domain adapted mid-scale models. For example, xOffense uses a fine-tuned Qwen3 32B model, and achieves 79.17% sub task completion rate and beating out much larger, general purpose frontier models on specific cybersecurity reasoning tasks [34]. MAPTA offers tool grounded execution and end to end exploit validation for web applications, solving 76.9% of challenges on major benchmarks at an average cost of $0.073 per successful attempt [30].

6.2 Threat Landscape for AI Security Systems

The threat landscape affecting AI security systems consists of both traditional cybersecurity threats and emerging AI-specific attack vectors. Cybercriminal organizations, insider threats, nation-state actors, and advanced persistent threat (APT) groups increasingly recognize the strategic importance of AI systems and actively seek opportunities to compromise them [35]. These adversaries may target the underlying infrastructure, manipulate sensor inputs, poison training data, exploit weaknesses in machine learning algorithms, or attack autonomous AI agents

responsible for threat detection and response [36]. Consequently, AI security systems must be protected not only against conventional cyberattacks but also against attacks specifically designed to influence model behavior and reduce detection capabilities [38].

Cybercriminal groups are increasingly adopting AI technologies to automate reconnaissance, phishing campaigns, malware development, vulnerability discovery, and attack execution [35]. At the same time, they attempt to undermine defensive AI systems by manipulating the information used during model training or inference. Adversarial attacks can introduce carefully crafted inputs that cause AI models to misclassify malicious activities as legitimate behavior or generate large numbers of false alarms [38]. Such attacks can significantly reduce the effectiveness of threat detection systems and create opportunities for successful intrusions. As AI-driven security operations become more common, the competition between offensive and defensive AI capabilities is expected to intensify [35].

Insider threats represent another significant risk to AI-based security systems. Employees, contractors, administrators, or trusted third parties may intentionally or unintentionally compromise AI systems by manipulating training datasets, modifying model parameters, bypassing security controls, or exposing sensitive information [36]. Because insiders often possess legitimate access privileges and detailed knowledge of system architectures, their actions can be particularly difficult to detect. Data poisoning attacks are among the most serious insider-related threats, as malicious modifications to training data may cause models to learn incorrect behaviors, reduce detection accuracy, or systematically ignore specific threats [37]. In critical infrastructure environments, such manipulation may directly affect operational safety and security.

Nation-state actors and advanced persistent threat groups pose some of the most sophisticated threats to AI-enabled security systems [40]. These actors typically possess significant technical capabilities, financial resources, and long-term strategic objectives. Their operations frequently target critical infrastructure sectors such as transportation, energy, telecommunications, defense, and maritime logistics [40]. Unlike conventional cybercriminals, nation-state actors often combine cyber operations with information influence campaigns, supply chain attacks, and physical sabotage. AI systems responsible for situational awareness and threat detection may become primary targets because compromising these systems can significantly reduce an organization's ability to detect and respond to attacks. Adversaries may therefore attempt to manipulate sensor data, corrupt AI models, interfere with communication channels, or exploit vulnerabilities in automated decision-support systems [41].

A particularly important category of threats involves adversarial machine learning attacks. These attacks exploit the mathematical properties of machine learning models to influence their outputs [39]. Common attack methods include adversarial examples, data poisoning, model extraction, model inversion, and evasion attacks. Adversarial examples involve small, carefully crafted modifications to input data that cause AI models to produce incorrect predictions while remaining nearly invisible to human observers [38]. Data poisoning attacks manipulate training datasets to influence model behavior [37], while model extraction attacks attempt to reconstruct proprietary machine learning models through repeated interactions [42]. Model inversion attacks seek to recover sensitive information from trained models, potentially exposing confidential operational

data [43]. Such attacks demonstrate that highly accurate AI systems can remain vulnerable to manipulation if security considerations are not integrated into their design [44].

The emergence of large language models and autonomous AI agents has further expanded the threat landscape. Modern AI agents can perform complex tasks, interact with external systems, access operational tools, and make decisions with limited human supervision. While these capabilities provide substantial benefits for automation and cyber defense, they also create new attack surfaces. Prompt injection attacks can manipulate agent behavior by introducing malicious instructions through user inputs or external data sources. Additional risks include unauthorized tool usage, privilege escalation, information leakage, and autonomous execution of harmful actions [45]. As AI agents increasingly participate in security operations, securing their behaviour and limiting their privileges become critical requirements.

The integration of AI into cyber-physical systems significantly broadens the attack surface by creating dependencies between digital systems and physical processes [41]. AI-enabled security platforms often rely on information from surveillance cameras, radar systems, thermal sensors, industrial control systems, communication networks, and environmental monitoring devices. Consequently, attackers can target both cyber and physical components simultaneously. Examples include sensor spoofing attacks that provide false information to AI systems, false data injection attacks against industrial processes, GPS and GNSS spoofing affecting navigation systems, manipulation of surveillance systems, and attacks against autonomous drones [46]. Such cyber-physical attacks may undermine situational awareness, disrupt operations, and cause AI systems to make incorrect decisions based on manipulated information.

Supply chain risks represent an additional concern for AI-enabled environments. Modern AI systems depend on a complex ecosystem of software libraries, cloud services, pretrained models, sensor manufacturers, communication providers, and third-party data sources. A compromise within any component of this ecosystem can affect the integrity and trustworthiness of the entire system. Supply chain attacks targeting AI development tools, machine learning frameworks, or external data providers have become increasingly common and may introduce vulnerabilities that remain undetected for extended periods [47].

The threat landscape is further complicated by the emergence of hybrid threats that combine cyber, physical, informational, economic, and social dimensions. Adversaries may use AI technologies to generate misinformation, automate influence operations, identify vulnerabilities, or coordinate attacks across multiple domains [35]. Simultaneously, they may target defensive AI systems to degrade situational awareness and disrupt decision-making processes. Such hybrid attacks are particularly relevant in critical infrastructure environments where disruptions can have significant societal, economic, and national security consequences [40].

6.3 Attack Vectors Against AI Systems

AI-based security systems are exposed to a wide range of attack vectors that target different stages of the AI lifecycle, including data collection, model training, model deployment, inference, and operational decision-making [36]. Unlike traditional information systems, AI systems rely heavily on large volumes of data and complex machine learning models, creating unique vulnerabilities that adversaries can exploit [48]. Attackers may seek to manipulate AI-generated decisions, reduce detection accuracy, bypass security controls, steal intellectual property, or

compromise operational trustworthiness [49]. Understanding these attack vectors is essential for developing resilient and secure AI systems capable of operating in hostile and rapidly evolving threat environments [50].

One of the most widely studied attack vectors is input manipulation through adversarial examples. Adversarial attacks involve carefully crafted modifications to input data that cause machine learning models to generate incorrect outputs while appearing normal to human observers [38]. In computer vision systems, small perturbations to images may cause an object detection model to misclassify targets or fail to detect them entirely [38]. Similarly, adversarial modifications to network traffic data may enable malicious activities to evade AI-based intrusion detection systems [51]. Because these attacks target the decision-making logic of machine learning models rather than exploiting software vulnerabilities, they can be difficult to detect using conventional cybersecurity controls [38].

Another significant attack vector is data poisoning during model training. Machine learning models depend on the quality and integrity of their training datasets. If attackers gain access to the data collection or training process, they may intentionally inject manipulated, mislabeled, or malicious data into the training set. Such poisoning attacks can alter model behavior, introduce hidden biases, create backdoors, or reduce overall detection performance. In critical infrastructure environments, poisoned models may fail to recognize specific threats, systematically ignore malicious activities, or generate misleading situational awareness information. Since poisoning attacks occur during the training phase, their effects may remain hidden until the model is deployed in operational environments [37].

Compromised sensors and data sources represent a particularly important threat in multi-modal AI systems. Modern anomaly detection and situational awareness platforms rely on information from a wide variety of sensors, including surveillance cameras, radar systems, thermal imaging devices, industrial sensors, access control systems, and communication networks. If attackers manipulate or compromise these data sources, the AI system may base its decisions on incorrect or fabricated information. Sensor spoofing attacks, false data injection attacks, manipulated video feeds, and compromised telemetry sources can significantly degrade the accuracy and reliability of AI-generated assessments. In cyber-physical systems, such attacks may directly affect operational safety and decision-making processes [52].

The communication infrastructure connecting sensors, AI models, and operational systems also presents numerous attack opportunities. Communication channel attacks, such as man-in-the-middle (MITM) attacks, eavesdropping, session hijacking, and data interception, target the transmission of information between system components. Attackers may intercept, modify, delay, or inject data while it is being transferred between sensors, processing platforms, cloud services, and end users. If communication channels are not adequately protected through encryption, authentication, and integrity verification mechanisms, adversaries may manipulate critical information without directly compromising the AI system itself. Such attacks can result in false situational awareness, incorrect threat assessments, and operational disruptions [52].

AI systems are also vulnerable to model deployment and infrastructure vulnerabilities. Once trained, machine learning models are typically deployed through application programming interfaces (APIs), cloud services, edge computing platforms, or integrated operational systems

[50]. Insecure APIs, weak authentication mechanisms, improper access controls, and software vulnerabilities can provide attackers with opportunities to manipulate model behavior or gain unauthorized access to AI services [45]. Attackers may exploit deployment weaknesses to extract model parameters, launch denial-of-service attacks, modify inference results, or access sensitive operational data. As organizations increasingly deploy AI capabilities through distributed and cloud-native architectures, securing the deployment environment becomes a critical component of AI security [50].

Beyond these primary attack vectors, AI systems may also be exposed to model extraction and model inversion attacks. Model extraction attacks attempt to reconstruct proprietary machine learning models by repeatedly querying deployed AI services and analyzing their outputs. Such attacks may reveal valuable intellectual property and enable adversaries to develop more effective evasion strategies. Model inversion attacks seek to recover sensitive information contained within training datasets, potentially exposing confidential operational, personal, or commercial data. These attacks highlight the importance of protecting both the AI model and the data used during its development [53].

The emergence of large language models and autonomous AI agents has introduced additional attack vectors, including prompt injection, unauthorized tool usage, context manipulation, and information leakage. Attackers may attempt to influence agent behavior through malicious prompts or manipulated external data sources, causing the AI system to perform unintended actions or reveal sensitive information. As AI agents become increasingly integrated into cybersecurity operations and critical infrastructure management, these threats require dedicated mitigation strategies [54].

The complexity of modern AI ecosystems means that attacks often target multiple components simultaneously. An attacker may compromise a sensor, manipulate transmitted data, exploit a vulnerable API, and ultimately influence the decisions generated by an AI model. Consequently, AI security must be approached from a holistic perspective that considers the entire data lifecycle, from data acquisition and transmission to model training, deployment, and operational use [50]. Protecting AI systems against these attack vectors requires a combination of technical, organizational, and operational controls. Effective defense mechanisms include secure data governance, adversarial training, robust model validation, sensor authentication, encrypted communications, access control mechanisms, continuous monitoring, and human oversight [50]. In critical infrastructure environments, where the consequences of AI failures can be severe, resilience against these attack vectors is essential for maintaining trustworthiness, operational continuity, and effective threat detection capabilities [52].

6.4 Risk Assessment

Risk assessment is a fundamental component of trustworthy and resilient AI systems. AI-enabled anomaly detection and security monitoring systems operate in dynamic and adversarial environments where both traditional cybersecurity threats and AI-specific attack vectors must be considered. Consequently, systematic risk assessment is essential for identifying vulnerabilities, prioritizing mitigation measures, and supporting informed decision-making throughout the lifecycle of AI systems [55].

Several internationally recognized frameworks provide structured approaches for risk assessment. Among the most widely adopted are the National Institute of Standards and Technology (NIST) Special Publication 800-30, Guide for Conducting Risk Assessments, and the ISO/IEC 27005 standard for information security risk management. These frameworks provide methodologies for identifying assets, threats, vulnerabilities, potential impacts, and risk treatment strategies. While originally developed for information security management, their principles are equally applicable to AI-enabled systems and cyber-physical environments. In the context of AI-based security systems, risk assessment extends beyond traditional cybersecurity concerns to include risks associated with machine learning models, training data, sensor infrastructures, autonomous decision-making processes, and human-AI interactions. A comprehensive assessment should therefore evaluate not only the probability of cyberattacks but also the trustworthiness, robustness, resilience, and operational reliability of AI components [57].

A key element of risk assessment is the evaluation of the likelihood of attack. Likelihood refers to the probability that a specific threat actor will successfully exploit a vulnerability within the AI system. Factors influencing likelihood include the attractiveness of the target, the capabilities and motivations of potential adversaries, the complexity of the attack, the accessibility of vulnerable components, and the effectiveness of existing security controls [55]. In AI environments, likelihood assessments should consider both conventional cyber threats and AI-specific attacks such as adversarial examples, data poisoning, model extraction, sensor spoofing, and prompt injection [58]. Systems deployed in critical infrastructure sectors may face elevated threat levels due to their strategic importance and exposure to nation-state adversaries [59].

The second major assessment dimension is the impact on system operations. Impact analysis evaluates the potential consequences of a successful attack or system failure [55]. In AI-enabled security systems, impacts may include degraded threat detection performance, loss of situational awareness, false alarms, missed detections, operational downtime, financial losses, safety incidents, reputational damage, or regulatory non-compliance [56]. In cyber-physical systems, impacts may extend beyond digital environments and affect physical operations, industrial processes, transportation systems, or critical services. Risk assessments should therefore consider both direct and indirect consequences as well as cascading effects across interconnected systems and infrastructures [61].

The third key component is the evaluation of detectability and response capability. Not all attacks have equal operational significance; risks are influenced not only by the probability and impact of an attack but also by an organization's ability to detect, contain, and recover from the event. AI systems should be assessed based on their ability to identify anomalous behavior, recognize attacks against the AI itself, generate actionable alerts, and support rapid incident response. Organizations with mature monitoring capabilities, incident response procedures, backup mechanisms, and recovery processes are generally better positioned to mitigate the effects of successful attacks. Conversely, attacks that remain undetected for extended periods often present the greatest risk because they allow adversaries to operate without interruption [62].

For AI-based systems, risk assessment should also consider model-specific factors such as training data quality, model robustness, explainability, uncertainty estimation, and resilience to degraded operating conditions. Multi-modal AI systems introduce additional dependencies that

must be evaluated, including sensor reliability, data fusion mechanisms, communication infrastructures, and external data sources. A vulnerability in any component may affect the performance of the entire system and therefore contribute to overall risk [60].

A practical approach for AI risk assessment is to combine likelihood, impact, and detectability into a structured risk scoring model. High-risk scenarios typically involve attacks that are relatively easy to execute, have severe operational consequences, and are difficult to detect or mitigate. Examples include successful data poisoning attacks during model training, coordinated sensor spoofing campaigns, or compromises of centralized AI decision-support platforms. Conversely, attacks that are difficult to execute, have limited consequences, or can be rapidly detected and contained generally represent lower levels of risk [55].

Risk assessment should not be treated as a one-time activity. The threat landscape, operational environment, and AI technologies continuously evolve, creating new vulnerabilities and changing risk profiles over time. Continuous risk monitoring, periodic reassessment, security testing, red-team exercises, and model validation activities are therefore essential components of AI governance and operational resilience. Organizations should regularly review risk assumptions, evaluate emerging attack techniques, and update mitigation measures accordingly [56].

In critical infrastructure environments, effective risk assessment supports the development of resilient and trustworthy AI systems by enabling organizations to prioritize security investments, implement appropriate controls, and prepare for emerging threats. By systematically evaluating attack likelihood, operational impact, and detection and response capabilities, organizations can better understand their risk exposure and improve the overall security and resilience of AI-enabled security systems [56].

7 SENSOR PLATFORM ROBUSTNESS

7.1 Sensor Reliability and Redundancy

Sensor reliability is critical in CPS. Redundancy strategies include:

- Multiple sensors for the same variable,
- Diverse sensing modalities,
- Cross-validation between sensors.

A key aspect of ensuring the robustness and resilience of AI-based systems lies in the reliability and quality of the sensory input. In particular, the performance of AI-based perception systems is highly dependent on the visibility and detectability of relevant objects within the scene. In real-world deployment scenarios, environmental conditions such as low illumination, high dynamic range, adverse weather, and thermal variability can significantly degrade image quality, thereby impacting the performance of downstream AI algorithms.

To address these challenges, a part of the project focused on systematic improvements of the sensing pipelines. By increasing object visibility and detectability across sensing modalities:

- AI models receive higher-quality, more informative inputs;
- detection and classification performance is maintained across a wider range of environmental conditions; and
- sensitivity to noise, illumination changes, and sensor artifacts is reduced.

As a result, the overall AI-based system becomes more robust and resilient to real-world variability, contributing to reliable operation even in challenging scenarios. The following subsections provide additional details into the specific improvements brought to the thermal and RGB imaging pipelines.

7.2 Objective measurement of object detectability in thermal images

A key enabler for improving sensor performance is the definition of objective and task-relevant metrics for object detectability in thermal imagery. Conventional image quality indicators (e.g., signal-to-noise ratio) are not sufficient to characterize performance in detection-oriented applications. Therefore, the project introduces and applies advanced evaluation methodologies that better reflect operational requirements. These methods provide a quantitative basis for assessing and optimizing thermal image quality with respect to AI-based detection tasks.

7.3 Optimization of thermal image quality

Building on the above metrics, the project implements improvements in the thermal image processing chain aimed at enhancing image uniformity, reducing noise, and improving contrast. Key improvements include:

- Enhanced calibration procedures, ensuring improved correction of sensor non-uniformities and stable radiometric response
- Refined gain map estimation, reducing fixed-pattern noise and spatial artifacts that affect detectability, which are particularly detrimental in low-signal scenarios
- Dedicated noise reduction techniques, preserving relevant structural details while minimizing both spatial and temporal noise

These enhancements result in improved thermal image clarity and object-background contrast, supporting more reliable detection performance, particularly in low-signal or high-noise conditions.

7.4 Optimization of RGB image quality in low-light conditions

In parallel to thermal imaging, RGB sensors are optimized to improve performance under reduced illumination. Given the importance of RGB data for semantic understanding, particular emphasis is placed on enhancing visibility in low-light environments:

- Low-light image enhancement, where an adaptive exposure control increases brightness while maintaining color consistency
- Noise-aware processing, with specific sensor knowledge and characterization, we limit noise amplification typically associated with low-light enhancement while preserving important details and color information
- Dynamic range optimization, enabling better preservation of scene details in challenging lighting conditions

These measures significantly improve the visibility of objects and persons in RGB imagery, thereby complementing thermal sensing and enabling more robust multi-modal perception.

7.5 Optimization of radar modalities

Millimeter-wave (mmWave) radar provides an important complementary sensing modality for AI-based situational awareness and anomaly detection systems, particularly in environments where optical sensors may be degraded, unavailable, or restricted by privacy regulations. Unlike RGB and thermal cameras, radar sensors do not capture identifiable visual information, making them inherently privacy-preserving while still enabling reliable detection, tracking, and classification of objects and human activity. As a result, mmWave radar represents an effective redundancy mechanism within multi-modal sensing architectures and contributes significantly to the overall resilience and trustworthiness of AI-based security systems.

One of the key advantages of mmWave radar is its ability to operate reliably under environmental conditions that often degrade optical sensing performance. Rain, fog, snow, smoke, darkness, glare, and rapidly changing illumination conditions can significantly reduce the effectiveness of RGB and thermal imaging systems. In contrast, radar sensing remains largely unaffected by these environmental factors, enabling continuous monitoring and situational awareness in challenging operational environments. This capability is particularly important in critical infrastructure applications, maritime environments, industrial facilities, and perimeter protection systems where uninterrupted sensing is required for safety and security.

From a resilience perspective, mmWave radar serves as an independent sensing modality that can maintain operational capability when optical systems fail or become unavailable. In multi-modal AI architectures, radar data can be used either as a supplementary information source or as a fallback sensing mechanism during degraded operational modes. This redundancy improves system robustness by reducing dependence on any single sensor technology and enabling graceful degradation when sensor failures occur.

The operational robustness of the radar-based sensing system is further enhanced through the use of advanced signal processing and object detection methodologies. Rather than relying solely on conventional image-based analysis, the radar processing pipeline utilizes five-dimensional (5D) data clustering, which combines spatial position, velocity, motion characteristics, and signal

intensity information to identify and track objects within the monitored area. This multidimensional representation provides a richer understanding of object behavior and improves the separation of genuine targets from environmental clutter and false reflections.

A particularly important feature is the utilization of radar signal intensity information to distinguish human targets from multipath reflections, ghost targets, and other sources of interference. Multipath propagation is a common challenge in radar systems, especially in complex environments containing metallic structures, buildings, vehicles, and industrial equipment. These reflections may create false detections that could otherwise reduce situational awareness accuracy. By incorporating intensity-based filtering and clustering mechanisms into the processing chain, the system can effectively identify genuine human clusters while suppressing false detections caused by reflected signals.

The combination of 5D clustering and intensity-based filtering enables reliable human detection without requiring computationally intensive optical image analysis or deep learning models. This approach improves operational efficiency while simultaneously increasing system robustness. The resulting radar-based sensing capability provides accurate localization and tracking performance even in scenarios where visual identification is impossible or undesirable.

From a privacy and regulatory perspective, radar sensing offers significant advantages over camera-based technologies. Since radar measurements do not contain identifiable personal information such as facial features or visual appearance, radar-based monitoring can often be deployed in environments where privacy regulations restrict the use of conventional video surveillance systems. This characteristic makes mmWave radar particularly attractive for public spaces, transportation hubs, industrial facilities, and critical infrastructure environments where both security and privacy requirements must be balanced.

Within the SINTRA framework, mmWave radar contributes to trustworthy AI by providing a resilient, privacy-preserving, and environmentally robust sensing modality that complements RGB and thermal imaging systems. The integration of radar data into multi-modal AI architectures improves overall system reliability, enhances anomaly detection capabilities, and supports continuous situational awareness under adverse operating conditions. Consequently, mmWave radar represents a key enabling technology for developing fault-tolerant and trustworthy AI systems capable of maintaining operational effectiveness in complex and dynamic security environments.

8 SUMMARY

The findings presented in this report are directly relevant to the objectives of the SINTRA project, which aims to develop trustworthy, resilient, and fault-tolerant AI-based solutions for anomaly detection, threat identification, and protection of critical infrastructures. The SINTRA project operates in environments where cyber, physical, and hybrid threats continuously evolve, requiring advanced situational awareness capabilities and robust decision-support mechanisms. The research and technologies analyzed in this study provide important scientific and technical foundations for achieving these objectives.

A key objective of SINTRA is the development of multi-modal AI systems capable of integrating heterogeneous data sources into a unified threat analysis platform. The report highlights the growing transition from single-modality anomaly detection systems toward multi-modal architectures that combine information from video surveillance, thermal imaging, radar systems, industrial sensors, communication networks, and external intelligence sources. Such capabilities directly support SINTRA project's vision of creating comprehensive situational awareness through the fusion of diverse sensor modalities and operational data streams.

The report also addresses one of the most significant challenges faced by the SINTRA project which is the detection of complex hybrid threats. Modern threat actors increasingly combine cyber, physical, informational, and operational attack vectors, making traditional rule-based security systems insufficient for identifying sophisticated attack chains. Multi-modal AI provides enhanced capabilities for recognizing anomalous behavior patterns, correlating events across multiple domains, and supporting proactive threat identification. These capabilities are particularly relevant for SINTRA demonstrators operating in critical infrastructure environments such as seaports and airports.

Trustworthiness represents a fundamental requirement for all AI systems developed within SINTRA. The study emphasizes that trustworthy AI must be lawful, ethical, robust, explainable, secure, and resilient throughout its operational lifecycle. In security-critical environments, trustworthiness extends beyond ethical considerations to include operational reliability under adverse conditions. The report identifies resilience, robustness, explainability, human oversight, accountability, and continuous monitoring as key principles for trustworthy AI deployment. These principles align closely with SINTRA's objective of ensuring that AI-generated alerts and recommendations remain reliable and actionable even during cyberattacks, sensor failures, communication disruptions, or other degraded operating conditions.

Another important contribution to SINTRA is the analysis of threats targeting AI systems themselves. As AI increasingly becomes a core component of critical infrastructure protection, adversaries are expected to target machine learning models, training datasets, sensors, and communication channels. The report examines attack vectors such as adversarial examples, data poisoning, sensor manipulation, communication channel attacks, and model deployment vulnerabilities. Understanding these threats is essential for SINTRA, as the project seeks not only to use AI for threat detection but also to ensure that the AI systems themselves remain secure and trustworthy. The mitigation approaches discussed, including adversarial training, secure data pipelines, uncertainty-aware decision making, and fault-tolerant architectures, provide valuable design guidance for future SINTRA implementations.

Operational resilience is another major area where the report contributes directly to SINTRA objectives. The study highlights that detection accuracy alone is insufficient for critical infrastructure protection. AI systems must continue operating despite equipment failures, missing

sensor inputs, environmental disturbances, or deliberate attacks. The proposed mechanisms, including sensor redundancy, adaptive multi-modal fusion, graceful degradation strategies, and continuous validation, directly support SINTRA's goal of maintaining operational continuity in challenging and adversarial environments.

Particularly relevant to SINTRA is the role of mmWave radar as a resilient and privacy-preserving sensing modality. The report demonstrates that mmWave radar can provide reliable detection and tracking capabilities even when optical sensors such as RGB or thermal cameras are degraded by fog, rain, snow, smoke, darkness, or other environmental factors. Furthermore, radar does not capture personally identifiable visual information, making it inherently compatible with privacy protection requirements. As a result, mmWave radar offers an effective redundancy mechanism within SINTRA's multi-modal sensing architecture and supports trustworthy AI operation in situations where optical sensing is unavailable or restricted.

The report further supports SINTRA's objectives related to secure data governance, explainable AI, and human-centered decision support. Explainable AI techniques improve transparency and operator trust by providing insights into the reasoning behind AI-generated alerts and recommendations. This capability is essential for security personnel responsible for validating anomalies, assessing risks, and making operational decisions. Combined with continuous monitoring and human oversight, explainability strengthens confidence in AI-assisted decision-making and contributes to the overall trustworthiness of the system.

In summary, the concepts, technologies, and methodologies presented in this report form a significant scientific foundation for the SINTRA project. They provide practical guidance for designing secure, trustworthy, and resilient multi-modal AI systems capable of operating in complex critical infrastructure environments. The strongest areas of relevance include multi-modal sensor fusion, hybrid threat detection, AI robustness and resilience, adversarial protection, operational continuity, privacy-preserving sensing, and trustworthy AI governance. Together, these capabilities support SINTRA's mission of developing next-generation AI-enabled security solutions for critical infrastructure protection.

9 REFERENCES

1. Chalapathy, R., & Chawla, S. (2019). *Deep Learning for Anomaly Detection: A Survey*. *arXiv preprint arXiv:1901.03407*. https://arxiv.org/abs/1901.03407
2. Zhou, C., Paudel, S., Song, W., Wang, M., et al. (2020). Deep Learning-Based Anomaly Detection: A Survey.
3. Goh, J., Adepu, S., Tan, M. Y., & Lee, Z. S. (2017). *Anomaly Detection in Cyber Physical Systems Using Recurrent Neural Networks*. In *Proceedings of the 2017 IEEE 18th International Symposium on High Assurance Systems Engineering (HASE)* (pp. 140–145). IEEE. https://doi.org/10.1109/HASE.2017.36
4. Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019). *Multimodal Machine Learning: A Survey and Taxonomy*. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(2), 423–443. https://doi.org/10.1109/TPAMI.2018.2798607
5. Tsai, Y.-H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L.-P., & Salakhutdinov, R. (2019). *Multimodal Transformer for Unaligned Multimodal Language Sequences*. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)* (pp. 6558–6569). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1656
6. Ma, X., Wu, J., Xue, S., Yang, J., Zhou, C., Sheng, Q. Z., Xiong, H., & Akoglu, L. (2021). *A Comprehensive Survey on Graph Anomaly Detection with Deep Learning*. IEEE Transactions on Knowledge and Data Engineering. https://arxiv.org/abs/2106.07178
7. P. Wu et al., 'VadCLIP: Adapting Vision-Language Models for Weakly Supervised Video Anomaly Detection', AAAI, vol. 38, no. 6, pp. 6074–6082, Mar. 2024, doi: 10.1609/aaai.v38i6.28423.
8. L. Zanella, B. Liberatori, W. Menapace, F. Poiesi, Y. Wang, and E. Ricci, 'Delving into CLIP latent space for Video Anomaly Recognition', Computer Vision and Image Understanding, vol. 249, p. 104163, Dec. 2024, doi: 10.1016/j.cviu.2024.104163.
9. A. Costanzino, P. Z. Ramirez, G. Lisanti, and L. Di Stefano, 'Multimodal Industrial Anomaly Detection by Crossmodal Feature Mapping', in 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA: IEEE, Jun. 2024, pp. 17234–17243. doi: 10.1109/CVPR52733.2024.01631.
10. B. Miao et al., 'Robust Modality-incomplete Anomaly Detection: A Modality-instructive Framework with Benchmark', 2024, arXiv. doi: 10.48550/ARXIV.2410.01737.
11. J. Wu et al., 'Assess and Guide: Multi-modal Fake News Detection via Decision Uncertainty', in Proceedings of the 1st ACM Multimedia Workshop on Multi-modal Misinformation Governance in the Era of Foundation Models, Melbourne VIC Australia: ACM, Oct. 2024, pp. 37–44. doi: 10.1145/3689090.3689389.
12. S. AlMarri, M. I. Liaqat, M. Z. Zaheer, S. Nawaz, K. Nandakumar, and M. Schedl, 'RobustA: Robust Anomaly Detection in Multimodal Data', 2025, arXiv. doi: 10.48550/ARXIV.2511.07276.
13. Y. Wang, H. Fu, W. Zou, and J. Jia, 'MMCert: Provable Defense against Adversarial Attacks to Multi-modal Models', 2024, arXiv. doi: 10.48550/ARXIV.2403.19080.

14. S. Jiang, Y. Ma, J. Zhou, Y. Bian, Y. Wang, and M. Liu, 'Resilient Multimodal Industrial Surface Defect Detection with Uncertain Sensors Availability', 2025, arXiv. doi: 10.48550/ARXIV.2509.02962.
15. D. Bogdoll et al., 'AnoVox: A Benchmark for Multimodal Anomaly Detection in Autonomous Driving', vol. 15629, 2025, pp. 206–223. doi: 10.1007/978-3-031-91767-7_15.
16. Z. Liu, X. Wu, W. Li, L. Yang, and S. Wang, 'Rethinking Metrics and Benchmarks of Video Anomaly Detection', 2025, arXiv. doi: 10.48550/ARXIV.2505.19022.
17. P. Wu et al., 'AVadCLIP: Audio-Visual Collaboration for Robust Video Anomaly Detection', 2025, arXiv. doi: 10.48550/ARXIV.2504.04495.
18. B. Leporowski et al., 'Audio-Visual Dataset and Method for Anomaly Detection in Traffic Videos', 2023, arXiv. doi: 10.48550/ARXIV.2305.15084.
19. F. Wang et al., 'Rapid in-flight image quality check for UAV-enabled bridge inspection', ISPRS Journal of Photogrammetry and Remote Sensing, vol. 212, pp. 230–250, Jun. 2024, doi: 10.1016/j.isprsjprs.2024.05.008.
20. Y. Yang, K. Lee, B. Dariush, Y. Cao, and S.-Y. Lo, 'Follow the Rules: Reasoning for Video Anomaly Detection with Large Language Models', 2024, arXiv. doi: 10.48550/ARXIV.2407.10299.
21. H. Zhang et al., 'Holmes-VAD: Towards Unbiased and Explainable Video Anomaly Detection via Multi-modal LLM', 2024, arXiv. doi: 10.48550/ARXIV.2406.12235.
22. T. Adewumi, L. Alkhaled, N. Gurung, G. van Boven, and I. Pagliai, 'Fairness and Bias in Multimodal AI: A Survey', 2024, arXiv. doi: 10.48550/ARXIV.2406.19097.
23. H. Du et al., 'Uncovering what, why and How: A Comprehensive Benchmark for Causation Understanding of Video Anomaly', in 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA: IEEE, Jun. 2024, pp. 18793–18803. doi: 10.1109/CVPR52733.2024.01778.
24. European Commission. (2019). *Ethics Guidelines for Trustworthy AI*. High-Level Expert Group on Artificial Intelligence. Publications Office of the European Union. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
25. Cisco. (2026, March 23). Cisco reimagines security for the agentic workforce. Cisco Newsroom. https://newsroom.cisco.com/c/r/newsroom/en/us/a/y2026/m03/cisco-reimagines-security-for-the-agentic-workforce.html
26. Jawale, N. (2026, March 30). RSAC 2026 recap: Agentic AI is the new attack surface and the new front line of defence. Opcito. https://www.opcito.com/blogs/rsac-2026-recap
27. Eliyahu, T., et al. (2025). Comparing AI Agents to Cybersecurity Professionals in Real-World Penetration Testing. arXiv. https://arxiv.org/abs/2512.09882
28. Security Boulevard. (2025). For $18 an Hour Stanford's AI Agent Bested Most Human Pen Testers in Study. https://securityboulevard.com/2025/12/for-18-an-hour-stanfords-ai-agent-bested-most-human-pen-testers-in-study/
29. AppSec Santa. (2026). AI Pentesting Agents 2026: Landscape, Tools & Benchmarks. https://appsecsanta.com/research/ai-pentesting-agents-2026
30. Chen, Y., et al. (2025). Multi-Agent Penetration Testing AI for the Web. arXiv. https://arxiv.org/abs/2508.20816

31. Anthropic. (2026). Project Glasswing: Securing critical software for the AI era. https://www.anthropic.com/glasswing
32. Horizon3.ai. (2025). NodeZero Becomes First AI to Fully Solve GOAD. https://horizon3.ai/news/press-release/horizon3-ais-nodezero-becomes-first-ai-to-fully-solve-game-of-active-directory-goad/
33. Waisman, N. (2025). The road to Top 1: How XBOW did it. XBOW. https://xbow.com/blog/top-1-how-xbow-did-it
34. Wang, Z., et al. (2025). xOffense: An Autonomous Multi-Agent Framework for Penetration Testing with Domain-Adapted Large Language Models. arXiv. https://arxiv.org/html/2509.13021v2
35. Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., hÉigeartaigh, S. Ó., Beard, S., Belfield, H., Farquhar, S., & Amodei, D. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv:1802.07228. https://arxiv.org/abs/1802.07228
36. Barreno, M., Nelson, B., Joseph, A. D., & Tygar, J. D. (2010). The security of machine learning. Machine Learning, 81(2), 121–148. https://doi.org/10.1007/s10994-010-5188-5
37. Biggio, B., Nelson, B., & Laskov, P. (2012). Poisoning attacks against support vector machines. In Proceedings of the 29th International Conference on Machine Learning (ICML 2012) (pp. 1467–1474).
38. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In International Conference on Learning Representations (ICLR 2015).
39. Huang, L., Joseph, A. D., Nelson, B., Rubinstein, B. I. P., & Tygar, J. D. (2011). Adversarial machine learning. In Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence (AISec '11) (pp. 43–58). https://doi.org/10.1145/2046684.2046692
40. European Union Agency for Cybersecurity. (2024). ENISA Threat Landscape 2024.
41. Humayed, A., Lin, J., Li, F., & Luo, B. (2017). Cyber-physical systems security—A survey. IEEE Internet of Things Journal, 4(6), 1802–1831. https://doi.org/10.1109/JIOT.2017.2703172
42. Tramèr, F., Zhang, F., Juels, A., Reiter, M. K., & Ristenpart, T. (2016). Stealing machine learning models via prediction APIs. In 25th USENIX Security Symposium (pp. 601–618).
43. Fredrikson, M., Jha, S., & Ristenpart, T. (2015). Model inversion attacks that exploit confidence information and basic countermeasures. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS 2015) (pp. 1322–1333). https://doi.org/10.1145/2810103.2813677
44. Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition, 84, 317–331. https://doi.org/10.1016/j.patcog.2018.07.023
45. OWASP Foundation. (2025). OWASP Top 10 for Large Language Model Applications. OWASP Top 10 for LLM Applications
46. Mo, Y., & Sinopoli, B. (2010). False data injection attacks in control systems. In Proceedings of the First Workshop on Secure Control Systems
47. European Union Agency for Cybersecurity. (2021). ENISA Threat Landscape for Supply Chain Attacks.

48. NIST. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). National Institute of Standards and Technology.
49. Huang, L., Joseph, A. D., Nelson, B., Rubinstein, B. I. P., & Tygar, J. D. (2011). Adversarial Machine Learning. Proceedings of AISec 2011.
50. European Commission. (2019). Ethics Guidelines for Trustworthy AI.
51. Papernot, N., McDaniel, P., Tramèr, F., Goodfellow, I. J., Jha, S., Celik, Z. B., & Swami, A. (2017). Practical Black-Box Attacks against Machine Learning. AsiaCCS 2017.
52. Humayed, A., Lin, J., Li, F., & Luo, B. (2017). Cyber-Physical Systems Security—A Survey. IEEE Internet of Things Journal, 4(6), 1802–1831.
53. Fredrikson, M., Jha, S., & Ristenpart, T. (2015). Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. ACM CCS.
54. OWASP Foundation. (2025). Agentic AI Threats and Mitigations.
55. National Institute of Standards and Technology (NIST). (2012). Guide for Conducting Risk Assessments (NIST Special Publication 800-30 Revision 1). Gaithersburg, MD: NIST.
56. National Institute of Standards and Technology (NIST). (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1.
57. European Commission High-Level Expert Group on Artificial Intelligence. (2019). Ethics Guidelines for Trustworthy AI. European Commission.
58. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples. International Conference on Learning Representations (ICLR).
59. European Union Agency for Cybersecurity (ENISA). (2024). ENISA Threat Landscape 2024.
60. Humayed, A., Lin, J., Li, F., & Luo, B. (2017). Cyber-Physical Systems Security—A Survey. IEEE Internet of Things Journal, 4(6), 1802–1831.
61. Lee, E. A. (2008). Cyber Physical Systems: Design Challenges. 11th IEEE International Symposium on Object-Oriented Real-Time Distributed Computing.
62. ISO 22301:2019. Security and Resilience—Business Continuity Management Systems—Requirements. International Organization for Standardization.