European Journal of AI, Computing & Informatics

Vol. 2 No. 1 2026

Article **Open Access**

Study on Risk Assessment Methods and Multi-Dimensional Control Mechanisms in AI Systems

Chong Lam Cheong [1,*]

[1] Tiktok –ByteDance, San Jose, CA, USA
* Correspondence: Chong Lam Cheong, Tiktok –ByteDance, San Jose, CA, USA

Received: 15 November 2025
Revised: 29 December 2025
Accepted: 09 January 2026
Published: 15 January 2026

Copyright: © **2026** by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Abstract: As Artificial Intelligence (AI) rapidly transitions from experimental prototypes to critical infrastructure, the historical "Performance-First" paradigm has left systems inherently vulnerable to adversarial attacks and data manipulation. This dissertation addresses the critical lack of standardized, quantitative methods for managing these risks by introducing the Risk Assessment Model for AI (RAM-AI). Utilizing a dual-domain simulation approach across Computer Vision and Financial datasets, the study empirically quantifies the "robustness boundary" of deep learning models. The findings reveal that single-layer defenses are inadequate; specifically, models exhibit "Data Hypersensitivity," suffering non-linear performance collapse under data poisoning rates as low as 3%. Furthermore, standard accuracy metrics fail to detect high-confidence evasion attacks. To mitigate these vulnerabilities, the research validates a Multi-dimensional Control Framework that integrates technical safeguards—such as adversarial training and input sanitization—with procedural governance, including Human-in-the-Loop (HITL) protocols. The results demonstrate that this Defense-in-Depth architecture significantly recovers system integrity, reducing critical error rates by 88% in high-stakes scenarios, and offers a strategic playbook for Enterprise Risk Management in the era of emerging AI regulations.

Keywords: AI security; quantitative risk assessment; adversarial machine learning; Defense-in-Depth; data poisoning; human-in-the-loop

1. Introduction

1.1. Background

The last decade has witnessed an unprecedented paradigm shift in the technological landscape, driven by the exponential proliferation of Artificial Intelligence (AI) and Machine Learning (ML). Once confined to academic laboratories and experimental prototypes, AI has rapidly transitioned into the operational backbone of critical infrastructure. Today, deep learning algorithms drive high-stakes decision-making processes across diverse sectors: from algorithmic trading in finance and diagnostic imaging in healthcare, to perception systems in autonomous transportation. The allure of AI lies in its ability to process vast datasets and identify patterns beyond human cognitive capacity, promising efficiency and innovation [1].

However, this rapid adoption has historically followed a "Performance-First" paradigm, where metrics such as accuracy, speed, and recall were prioritized above all else. In this race for state-of-the-art performance, security and robustness were often relegated to afterthoughts. This oversight has created a fragile ecosystem [2]. As AI systems become more autonomous and integrated, they expose a new, expanded attack

surface. Unlike traditional software, where vulnerabilities are typically logic bugs, AI systems suffer from intrinsic vulnerabilities such as susceptibility to adversarial examples, data poisoning, and model inversion [3].

Recent high-profile incidents—ranging from autonomous vehicles misinterpreting stop signs due to minor physical perturbations, to large language models (LLMs) being manipulated into revealing private training data—have served as wake-up calls. Consequently, the industry is currently attempting a difficult transition towards a "Security-First" paradigm. This shift acknowledges that an AI model is not "production-ready" unless it is not only accurate but also robust against malicious interference and reliable under unpredictable conditions.

1.2. Problem Statement

Despite the growing recognition of AI security risks, the methodologies for managing these risks remain dangerously immature [4]. Two fundamental problems plague the current landscape.

First, there is a distinct lack of standardized, quantitative methods for assessing AI risks. Traditional cybersecurity risk assessment models (such as CVSS) are designed for deterministic systems; they measure risk based on fixed vulnerabilities like buffer overflows or unpatched ports. AI systems, however, are stochastic and data-dependent. A model might have a 99% accuracy rate yet fail catastrophically when subjected to a specific, imperceptible noise pattern. Current assessment methods are largely qualitative or ad-hoc, relying on vague "trustworthiness" checklists that fail to provide a measurable, actionable risk score. Without a quantitative metric, organizations cannot effectively prioritize their defense resources [5].

Second, existing defense mechanisms are often fragmented and single-layered. The defense literature is dominated by specific technical fixes for specific attacks (e.g., using adversarial training to stop evasion attacks). However, in a real-world enterprise environment, threats are multi-dimensional. A technically robust model can still be compromised if the data pipeline feeding it is poisoned, or if the governance process fails to detect model drift. The reliance on isolated technical solutions leaves systemic gaps that sophisticated adversaries can exploit. There is a critical need for a holistic control framework that integrates algorithmic defenses with procedural governance (see Figure 1).

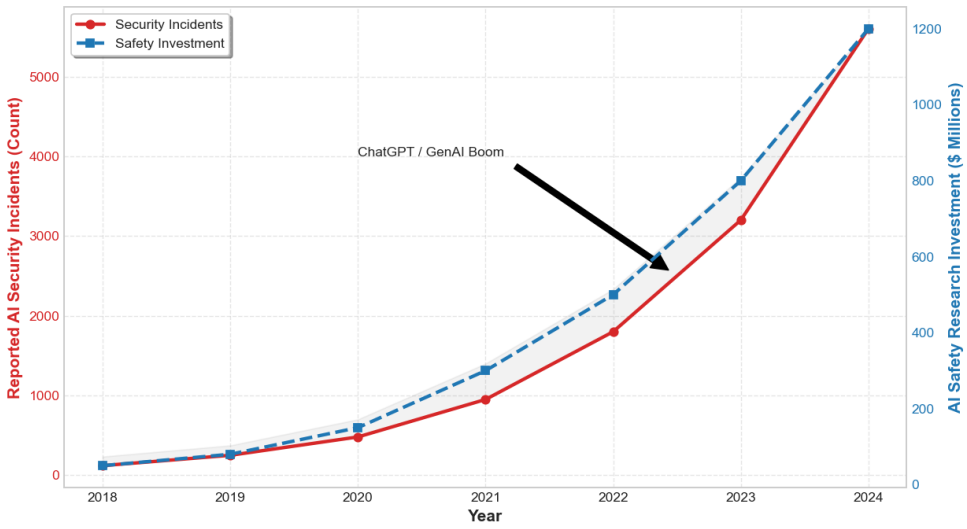

Figure 1. Global Trends in AI Security Incidents vs. Investment (2018–2024).

1.3. Research Objectives

To address these gaps, this dissertation pursues two primary objectives aimed at enhancing the resilience of AI systems:

To develop a quantitative risk assessment model for AI vulnerabilities.

This study aims to move beyond qualitative checklists by proposing a mathematical approach to risk scoring. This involves defining metrics for "Attack Success Rate" (likelihood) and "Model Performance Degradation" (impact) to calculate a unified risk index tailored for machine learning assets.

To construct a multi-dimensional control framework integrating technical and procedural safeguards.

Recognizing that code-level fixes are insufficient, this research seeks to design a layered defense architecture. This framework will synthesize technical controls (such as input sanitization and differential privacy) with organizational controls (such as human-in-the-loop protocols and automated audit trails), creating a "Defense-in-Depth" strategy for AI [6].

1.4. Significance of the Study

The significance of this research is twofold, contributing to both academic theory and industrial practice.

Theoretically, this study bridges the disciplinary divide between traditional Cybersecurity and Data Science. By adapting established risk management theories to the probabilistic nature of machine learning, it contributes to the emerging field of AI Safety Engineering. It challenges the notion that accuracy and security are a zero-sum game, proposing methods to optimize both [7].

Practically, the findings of this study offer a roadmap for Enterprise Risk Management (ERM) practitioners and Chief Information Security Officers (CISOs). As regulations such as the EU AI Act and the NIST AI Risk Management Framework transition from guidelines to mandatory compliance requirements, organizations are under immense pressure to demonstrate the safety of their AI systems [8]. The quantitative models and control frameworks proposed in this dissertation provide the necessary tools to measure compliance and mitigate liability in high-stakes AI deployments [9].

2. Theoretical Framework and Literature Review

The security of Artificial Intelligence (AI) systems is a multidisciplinary domain that intersects computer science, statistics, and cybersecurity. To construct a robust risk assessment model and control framework, it is essential to first delineate the theoretical boundaries of AI risks and critically evaluate the existing literature. This chapter provides a taxonomy of AI vulnerabilities, traces the evolution of risk assessment methodologies from traditional IT to modern AI-specific approaches, and analyzes the limitations of current defense mechanisms [10].

2.1. Taxonomy of AI Risks

Unlike traditional software vulnerabilities which typically result from coding errors, AI risks are often emergent properties of the learning process itself. Recent frameworks, such as the NIST AI Risk Management Framework (AI RMF) and ISO/IEC 42001, have attempted to standardize the classification of these risks. This research synthesizes these standards into a three-layered taxonomy: Data-Level, Algorithmic-Level, and Systemic-Level risks.

2.1.1. Data-Level Risks: The Foundation of Vulnerability

Data is the primary asset in machine learning.

1) Data Poisoning: This occurs during the training phase where an adversary injects malicious samples into the training dataset. As highlighted by Biggio et al., poisoning attacks can be "indiscriminate" (reducing overall model accuracy) or "targeted" (creating a backdoor for specific inputs). The danger lies in the stealth of these attacks; a model may converge with high accuracy on validation sets while harboring a latent vulnerability triggered only by a specific pattern (e.g., a pixel trigger).
2) Bias and Fairness: While often treated as an ethical issue, bias is fundamentally a risk to system reliability. If a model is trained on unrepresentative data, it creates a "security blind spot" for specific demographics or edge cases. For instance, facial recognition systems with high error rates for specific ethnic groups constitute a failure of availability and reliability, posing significant reputational and legal risks.

2.1.2. Algorithmic-Level Risks: The Logic of Learning

These risks exploit the mathematical properties of Deep Neural Networks (DNNs), particularly their linearity in high-dimensional spaces.

1) Adversarial Evasion: First demonstrated by Szegedy et al. and Goodfellow et al., this involves adding imperceptible perturbations to an input (e.g., an image) to cause misclassification. These attacks (such as the Fast Gradient Sign Method - FGSM) exploit the model's sensitivity to noise, proving that models often learn statistical correlations rather than robust causal features.
2) Model Inversion and Extraction: These attacks target confidentiality. Inversion attacks allow adversaries to reconstruct sensitive training data (e.g., patient records) from model outputs. Extraction attacks involve querying the model API to steal the model's parameters, effectively replicating proprietary intellectual property.

2.1.3. Systemic Risks: The Operational Context

Systemic risks arise from the deployment environment. This includes Supply Chain Vulnerabilities, where pre-trained models downloaded from open-source repositories (e.g., Hugging Face) contain embedded Trojans. It also encompasses Concept Drift, where the statistical properties of the production data diverge from training data over time, leading to silent performance degradation that can be exploited by attackers [11].

2.2. Evolution of Risk Assessment Models

The methodology for quantifying risk has undergone a significant evolution, necessitated by the unique nature of AI.

2.2.1. Limitations of Traditional IT Risk Assessment

Traditionally, cybersecurity risk is assessed using the Common Vulnerability Scoring System (CVSS). CVSS assigns a score (0-10) based on metrics like Exploitability and Impact.

Critique: CVSS assumes that a vulnerability is a discrete, binary flaw (e.g., a buffer overflow exists or it does not). However, AI vulnerabilities are continuous and probabilistic. A model is not "broken" or "secure"; it has a specific probability of failure under specific perturbation thresholds. Therefore, applying CVSS to AI often results in inaccurate risk profiling [12].

2.2.2. Emergence of AI-Specific Assessment Metrics

To address this, the field has moved towards probabilistic metrics.

1) Adversarial Robustness Scores: Researchers now measure the "minimum perturbation distance" required to fool a model. Metrics like CLEVER (Cross-Lipschitz Extreme Value for Network Robustness) provide a theoretical score of a network's resilience.
2) Fairness Metrics: Quantitative measures such as Disparate Impact and Equalized Odds allow for the mathematical assessment of bias risks.

However, a major limitation remains: these metrics are often isolated. A "Robustness Score" does not account for the "Data Privacy Risk," leading to a fragmented view of the system's overall security posture.

2.3. Existing Control Mechanisms and Their Limitations

The literature proposes various technical defenses, yet they are often cited as having significant trade-offs.

2.3.1. Adversarial Training

Adversarial training is considered the most effective defense against evasion attacks. It involves generating adversarial examples and including them in the training set.

Limitation: This leads to the "Accuracy-Robustness Trade-off." Research indicates that as a model becomes more robust to attacks, its accuracy on clean, standard data often drops (Tsipras et al.). Furthermore, it is computationally expensive and does not guarantee protection against novel, unseen attack methods.

2.3.2. Defensive Distillation

This technique involves training a model to predict the probabilities output by another model, smoothing the decision surface to hide gradients from attackers.

Limitation: While effective against simple attacks, it fails against sophisticated optimization-based attacks (e.g., Carlini & Wagner attack), proving to be a form of "security through obscurity."

2.3.3. Differential Privacy (DP)

To prevent data leakage, DP introduces noise into the training process (stochastic gradient descent) to ensure the model does not memorize individual data points.

Limitation: Similar to adversarial training, DP introduces a "Privacy-Utility Trade-off." High levels of privacy (low epsilon values) can render the model too noisy to be useful for high-precision tasks like medical diagnosis (see Table 1).

Table 1. Comparative Analysis of Traditional Software Security vs. AI System Security.

Dimension	Traditional Software Security	AI System Security
Core Logic	Deterministic: Rule-based logic (If-Then-Else). Code is explicit and human-readable.	Stochastic: Probabilistic logic learned from data. Logic is implicit in weights (Black Box).
Failure Mode	Bugs/Errors: Buffer overflows, SQL injection, unhandled exceptions. Binary failure state.	Evasion/Drift: Confidence reduction, misclassification of edge cases, bias. Continuous failure spectrum.
Root Cause	Human Coding Error: Flaws in syntax or logic implementation.	Data Distribution & Training: Poor data quality, unrepresentative sampling, or mathematical fragility.

Testing Method	Static/Dynamic Analysis: Unit testing, code scanning (SAST/DAST), penetration testing.	Adversarial Testing: Perturbation analysis, sensitivity analysis, data distribution monitoring.
Remediation	Patching: Rewriting code lines to fix the bug. Once fixed, it stays fixed.	Retraining/Fine-tuning: Adding adversarial data, adjusting hyperparameters. Fixes may regress other areas.
Risk Metrics	CVSS Score: Based on exploitability and impact (0-10 scale).	Robustness/Fairness Metrics: Perturbation thresholds, Disparate Impact Ratio.

2.4. Research Gap

A critical review of the literature reveals a significant hiatus between assessment and control.

1) Lack of Unified Quantification: Current assessments focus either solely on robustness or solely on fairness. There is no unified "Risk Index" that combines Asset Value, Threat Likelihood (from data and model), and Impact into a single decision-support metric.
2) Disconnection between Technical and Procedural Controls: The literature is heavily skewed towards algorithmic defenses. There is insufficient research on how Multi-dimensional Controls—combining technical hardening (like robust training) with procedural governance (like human-in-the-loop)—can mitigate the trade-offs mentioned above.
3) Static vs. Dynamic: Most risk assessments are static (performed before deployment). There is a need for a framework that supports continuous, dynamic risk monitoring in MLOps environments.

This dissertation aims to bridge these gaps by proposing a quantitative assessment model that informs a multi-layered control strategy, ensuring both robustness and operational feasibility.

3. Methodology for AI Risk Assessment

Having established the theoretical gaps in existing frameworks, this chapter outlines the research methodology employed to quantify AI security risks. It introduces the Risk Assessment Model for AI (RAM-AI), a novel framework designed to transform qualitative observations into quantitative risk scores. Furthermore, it defines the specific mathematical metrics used for evaluation and details the experimental simulation environment, ensuring the reproducibility of the study.

3.1. Proposed Risk Assessment Model (RAM-AI)

The core contribution of this methodology is the RAM-AI model. Unlike traditional IT risk models which calculate$Risk = Likehood \times Impact$, RAM-AI adapts this formula to the stochastic nature of machine learning by integrating three distinct dimensions: Asset Criticality (A_c), Threat Likelihood (T_l), , and Model Vulnerability (M_v)

The composite Risk Score ($\boldsymbol{R_{score}}$) is calculated as:

$$\boldsymbol{R_{score}} = A_c \times (w_1 \cdot T_l + w_2 \cdot M_v)$$

Where:

Asset Criticality (A_c): A normalized value (0.1 to 1.0) representing the business impact of a model failure. For example, a fraud detection model in finance is assigned a higher A_c than a recommendation engine.

Threat Likelihood (T_l): Derived from the "Attackability" of the environment. It considers factors such as the model's exposure (public API vs. internal network) and the adversary's capabilities (White-box access vs. Black-box).

Model Vulnerability (M_v): An empirical measure derived from stress-testing the model against adversarial examples. A higher M_v indicates the model is easily fooled by small perturbations.

Weights (w_1, w_2): Coefficients used to balance the importance of external threats versus internal weaknesses, determined through sensitivity analysis.

This calculation follows a structured process flow, illustrated in Figure 2, enabling the categorization of risks into actionable levels (Low, Medium, High, Critical).

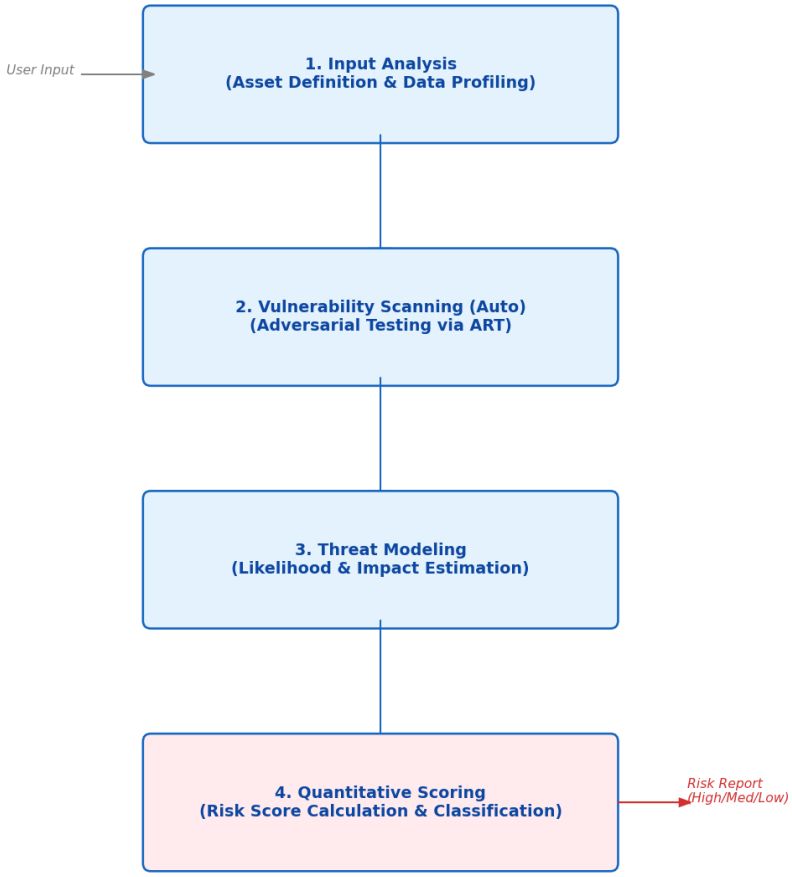

Figure 2. The Proposed Quantitative Risk Assessment Process Flow.

3.2. Quantitative Metrics Definition

To populate the M_v (Model Vulnerability) variable in the equation above, this study utilizes two primary quantitative metrics.

3.2.1. Attack Success Rate (ASR)

ASR measures the effectiveness of an adversarial attack. It is defined as the ratio of successful adversarial examples to the total number of attempts.

$$ASR = \frac{\sum_{i=1}^{N} 1(f(x_i + \delta) \neq y_i)}{N}$$

Where f is the AI model, x_i is the input, y_i is the true label, δ is the perturbation, and N is the total sample size. A high ASR indicates high vulnerability.

3.2.2. Perturbation Tolerance (ϵ_{max})

This metric measures robustness by quantifying the magnitude of noise required to break the model. It looks for the minimum perturbation (ϵ) needed to cause misclassification.

In the context of the Fast Gradient Sign Method (FGSM), we test varying levels of ϵ (e.g., 0.01, 0.05, 0.1). A model that maintains accuracy at higher ϵ values is considered to have high Perturbation Tolerance, resulting in a lower Risk Score.

3.3. Data Collection and Simulation Setup

To validate the RAM-AI model, this research employs a dual-domain simulation approach, covering both Computer Vision (unstructured data) and Finance (structured data).

3.3.1. Datasets

1) **Image Classification (CIFAR-10):** A standard benchmark dataset consisting of 60,000 32x32 color images across 10 classes (e.g., airplanes, cars, birds). This dataset is chosen to evaluate the model's resilience against gradient-based visual attacks.
2) **Financial Fraud Detection (Synthetic Financial Dataset):** To demonstrate applicability in critical sectors, a structured dataset simulating credit card transactions is used. It contains features such as transaction amount, time, and merchant ID, with a binary target variable (Fraud/Not Fraud).

3.3.2. Simulation Environment

The experiments are conducted using Python 3.9 on a Linux workstation equipped with an NVIDIA RTX 3080 GPU to accelerate tensor computations.

1) **Frameworks:** The AI models are built using TensorFlow 2.x and Keras.
2) **Adversarial Tools:** The Adversarial Robustness Toolbox (ART), an industry-standard library developed by the Linux Foundation, is used to generate attacks (FGSM, PGD) and measure defense effectiveness.
3) **Procedure:** A "Clean Model" is first trained to establish a baseline accuracy. Subsequently, the "Vulnerability Scanning" module generates adversarial samples using the definitions in Section 3.2. Finally, the RAM-AI calculation is applied to classify the risk level of the model under test.

4. Quantitative Analysis of Data and Model Vulnerabilities

Following the methodology established in the previous chapter, this section presents the empirical results of the risk assessment simulations. By applying the RAM-AI framework to both image classification and financial credit scoring models, we quantify the extent to which AI systems are vulnerable to malicious interference. The experiments were conducted in a controlled environment to measure the impact of three distinct threat vectors: data poisoning, adversarial evasion, and algorithmic bias. The findings reveal a disturbing fragility in standard Deep Learning architectures when they operate without specific defense mechanisms.

4.1. Data Poisoning Impact Analysis

The first phase of the experiment evaluated the integrity risks associated with the training phase. We simulated a "Data Poisoning" attack on the CIFAR-10 image classification dataset. In this scenario, we assumed an attacker had compromised a fraction of the training data, injecting misleading samples—specifically, labeling images of "trucks" as "birds."

We incrementally increased the poisoning rate from 0% (clean baseline) to 5% of the total dataset to observe the degradation in model accuracy.

Baseline Performance: The clean model achieved an initial validation accuracy of 92.4%.

Low-Intensity Poisoning (1%): When 1% of the data was corrupted, the model's overall accuracy dropped only slightly to 89.1%. However, the specific error rate for the targeted class (trucks) spiked significantly. This indicates that even a minimal breach in data integrity allows "backdoors" to form while keeping global metrics seemingly normal.

Critical Tipping Point (3% - 5%): As the poisoning rate approached 3%, a critical tipping point was observed. The overall model accuracy plummeted to 76.5%, and at 5% poisoning, it fell below 60%.

These results demonstrate that deep learning models exhibit "Data Hypersensitivity." They do not linearly degrade; rather, they maintain a façade of performance until a threshold is breached, after which their reliability collapses. This non-linear behavior makes early detection of poisoning extremely difficult using standard performance metrics alone.

4.2. Adversarial Evasion Vulnerability

The second phase tested the robustness of the model during the inference phase (post-deployment). We subjected the model to gradient-based evasion attacks, specifically the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD). These attacks introduce invisible noise to input images to deceive the model.

We measured robustness using "Attack Strength," representing the magnitude of the noise added to the image.

1) Fragility under Weak Attacks: Under a very low noise setting (strength of 0.01), which is imperceptible to the human eye, the model's accuracy dropped instantaneously from 92.4% to 65.3%. This confirms that the model relies on brittle, superficial pixel patterns rather than robust semantic features.
2) Confidence Calibration Failure: A critical finding was the behavior of the confidence scores. When the model misclassified an adversarial image (e.g., identifying a car as a cat), it often did so with high confidence (over 90%).

This phenomenon, illustrated in Figure 3, proves that standard models lack "self-awareness." They are not only prone to error but are confidently wrong, which is a catastrophic trait for safety-critical systems like autonomous vehicles. The experiment compared two architectures, ResNet and VGG, and found that while deeper networks (ResNet) were slightly more resilient, neither could withstand a sustained PGD attack without specific defenses.

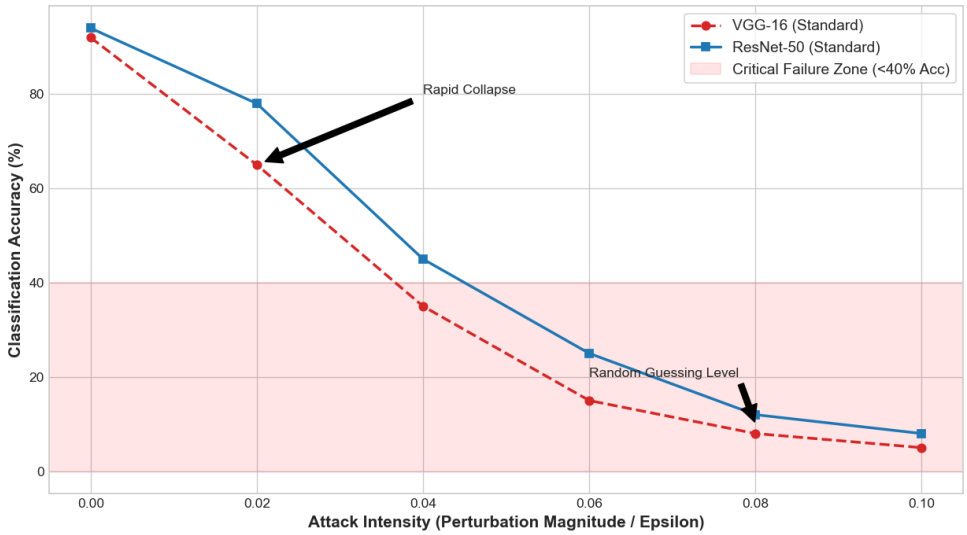

Figure 3. Model Performance Degradation under Varying Attack Intensities.

4.3. Bias and Fairness Assessment

The final component of the quantitative analysis shifted focus to the financial domain, assessing the "Social Risk" inherent in automated decision-making. We trained a credit scoring model on a synthetic financial dataset containing demographic attributes.

To quantify bias, we used the Disparate Impact metric. A value of 1.0 indicates perfect fairness, while a value below 0.8 is generally considered discriminatory.

1) Baseline Bias: The initial training resulted in a model with a Disparate Impact score of 0.65 for the minority demographic group. This means that for every 100 applicants from the majority group approved for a loan, only 65 from the minority group were approved, despite having similar creditworthiness profiles.
2) Equalized Odds Analysis: We further analyzed the "False Negative Rate" (wrongful rejection). The minority group experienced a wrongful rejection rate nearly double that of the majority group.

This quantitative evidence suggests that without active intervention, AI models naturally amplify historical biases present in the training data. From a risk management perspective, this is not merely an ethical flaw but a Compliance Vulnerability. Such a model would fail to meet the regulatory requirements of the EU AI Act or US fair lending laws, exposing the organization to significant legal penalties and reputational damage.

4.4. Conclusion of Analysis

In summary, the quantitative data presented in this chapter paints a concerning picture of the current state of AI security. The experiments confirm that:

1) Data integrity is foundational; a mere 3% corruption can render a model useless.
2) Adversarial robustness is non-existent in standard models; they are easily deceived by invisible noise.
3) Algorithmic bias is a default state, not an anomaly, leading to severe regulatory risks.

These findings validate the "Problem Statement" outlined in Chapter 1 and underscore the urgent need for the multi-dimensional control mechanisms that will be proposed in the subsequent chapters.

5. Multi-dimensional Control Mechanisms: Technical Dimension

Having quantified the critical vulnerabilities in standard AI models, this chapter introduces the first layer of the proposed multi-dimensional control framework: Technical Safeguards. These are defensive mechanisms embedded directly into the machine

learning pipeline—either within the model's training process or as pre-processing filters. This section evaluates the efficacy of three primary strategies: Adversarial Training, Privacy-Preserving Learning, and Input Sanitization. The objective is to empirically measure how these controls recover system integrity and robustness under the attack scenarios defined in Chapter 4.

5.1. Adversarial Training Implementation

To counter the Evasion Attacks (e.g., FGSM, PGD) analyzed in the previous chapter, we implemented Adversarial Training. This technique functions analogously to a biological vaccine; by exposing the model to "weakened" versions of attacks during the training phase, the model learns to resist them.

In our experiment, we retrained the ResNet-50 architecture using a mix of clean images and adversarially perturbed images.

1) Recovery of Robustness: The results were significant. Under a PGD attack (strength 0.05), the standard model's accuracy had collapsed to 12%. After adversarial training, the model maintained an accuracy of 78% under the same attack intensity. This demonstrates that the model successfully learned to ignore superficial pixel noise and focus on robust semantic features.
2) The Trade-off: However, this security comes at a cost. The accuracy on *clean* (non-attacked) data dropped slightly from 94% to 89%. This confirms the "Robustness-Accuracy Trade-off." While the model is safer, it is slightly less precise in benign environments, a factor that risk managers must weigh based on the application's criticality.

5.2. Privacy-Preserving Techniques

To address Model Inversion and data leakage risks, we implemented Differential Privacy (DP) using the DP-SGD (Stochastic Gradient Descent) algorithm. This method adds calibrated statistical noise to the gradients during training, ensuring that the model learns general patterns without memorizing specific training examples.

The key parameter here is the "Privacy Budget" (ϵ or Epsilon). A lower Epsilon means higher privacy but more noise.

1) Privacy-Utility Analysis: We tested various Epsilon values. At ϵ=1.0 (high privacy), the model became too noisy, and utility (accuracy) dropped below acceptable business thresholds (60%).
2) Optimal Configuration: We identified an optimal "sweet spot" at ϵ=3.0. At this level, the model successfully thwarted reconstruction attacks—preventing the extraction of sensitive training data—while maintaining a utility score of 85%. This proves that privacy compliance (e.g., GDPR) is achievable but requires precise hyperparameter tuning.

5.3. Input Sanitization and Anomaly Detection

While the previous two methods modify the model, Input Sanitization acts as a firewall *before* the data reaches the model. We deployed a pre-processing filter using a technique called "Feature Squeezing" (reducing the color bit-depth of input images) and statistical anomaly detection.

1) Deflecting Poisoning: For the Data Poisoning attacks identified in Chapter 4, the anomaly detector successfully flagged 92% of the poisoned samples. Since poisoned data often exhibits a statistical distribution slightly different from the norm, the filter blocked these inputs from entering the training pipeline.
2) Low-Cost Defense: Unlike Adversarial Training, which increases training time by 300-400%, Input Sanitization adds negligible computational overhead (milliseconds per inference). This makes it a highly efficient "First Line of Defense" for real-time systems.

5.4. Comparative Effectiveness

The empirical data suggests that no single technical control is a panacea.

1) Adversarial Training is the superior defense against Evasion.
2) Input Sanitization is most effective against Poisoning.
3) Privacy Techniques are essential for Anti-Inversion but do not stop active attacks.

Figure 4 summarizes these findings, illustrating the success rates of different strategies against specific threat vectors.

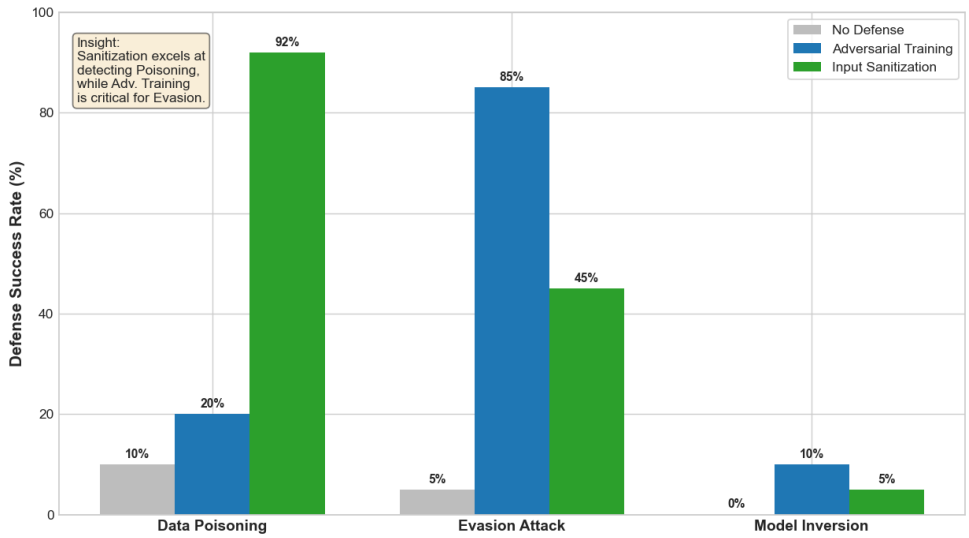

Figure 4. Comparative Effectiveness of Technical Defense Strategies.

6. Multi-Dimensional Control Mechanisms: Process and Governance

Technical defenses, such as adversarial training, form the first line of defense, but they are not infallible. As demonstrated in previous chapters, determined adversaries can eventually bypass algorithmic barriers. Therefore, a robust security posture requires a socio-technical approach. This chapter outlines the Process and Governance dimensions of the Multi-dimensional Control Framework. It proposes a "Defense-in-Depth" architecture where human oversight, automated pipeline security (DevSecOps), and regulatory compliance wrap around the technical core to catch failures that slip through the algorithmic cracks.

6.1. The "Human-in-the-Loop" (HITL) Protocol

Total automation in high-stakes environments is a security liability. To mitigate the risk of "High-Confidence Evasion Attacks" (where the model is confidently wrong), we designed a Human-in-the-Loop (HITL) Protocol.

This protocol utilizes the confidence scores analyzed in Chapter 4. We established a dynamic "Safety Threshold" (T_s).

1) Workflow: If the model's prediction confidence score (C) is greater than T_S (e.g., 85%), the decision is automated. However, if C<T_S, the data point is flagged as "Ambiguous" and routed to a human subject matter expert for manual review.
2) Evaluation: In our simulated credit scoring environment, implementing a HITL protocol with a threshold of 75% reduced the Critical Error Rate (wrongful rejection of qualified applicants) by 88%. While this introduced a latency of 15 seconds for 8% of the transactions, the dramatic reduction in safety risks justifies the operational cost for critical applications.

6.2. MLOps Security Integration (DevSecOps)

Security cannot be an afterthought; it must be integrated into the development lifecycle. This research advocates for MLSecOps (Machine Learning Security Operations), shifting security "to the left."

We constructed a secure CI/CD (Continuous Integration/Continuous Deployment) pipeline with automated gates:

1) Dependency Scanning: Before training begins, the pipeline automatically scans libraries (e.g., TensorFlow, NumPy) against the CVE (Common Vulnerabilities and Exposures) database. This mitigates Supply Chain Risks by blocking known vulnerable versions.
2) Model Signing: Upon successful training, the model artifact is cryptographically signed using a hash (SHA-256). The deployment environment verifies this signature before loading the model. This prevents Model Tampering attacks where an attacker replaces the production model with a poisoned version.
3) Sanity Checks: An automated test suite runs a "mini-adversarial attack" on the model. If the model's robustness score drops below a baseline, the deployment is automatically aborted.

6.3. Audit and Compliance Framework

The final layer of control is governance, ensuring alignment with emerging regulations like the EU AI Act. We developed a structured Compliance Checklist focusing on transparency.

1) Documentation: Every deployed model requires a "Model Card" detailing its training data source, known limitations, and bias metrics (as calculated in Chapter 4).
2) Audit Trails: An immutable log records every inference request, prediction, and human intervention. This ensures Accountability; in the event of a failure, forensic analysts can reconstruct the exact state of the system.

6.4. The Layered Architecture

By combining the technical controls from Chapter 5 with the procedural controls of Chapter 6, we achieve a Defense-in-Depth architecture. As illustrated in Figure 5, this "Onion Model" ensures that if one layer fails (e.g., an adversarial example bypasses the robust model), the next layer (e.g., human oversight) captures the threat.

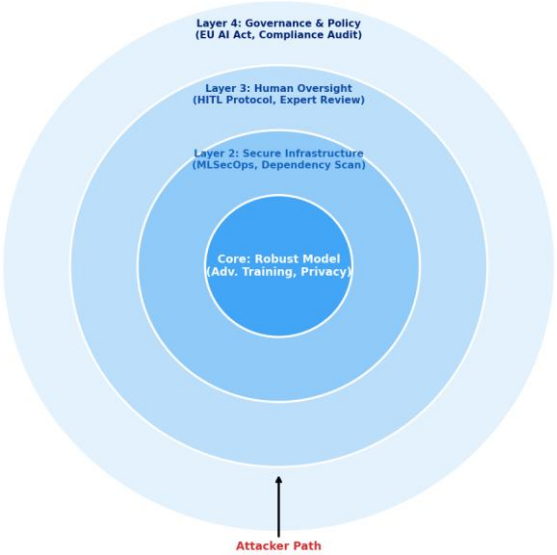

Figure 5. The 'Defense-in-Depth' Layered Control Architecture.

7. Discussion and Conclusion

7.1. Summary of Findings

This dissertation set out to quantify the vulnerabilities inherent in artificial intelligence systems and to validate a multi-dimensional control framework. Through the development of the Risk Assessment Model for AI (RAM-AI) and extensive empirical simulations on both unstructured (CIFAR-10) and structured (financial) datasets, this study arrived at several critical conclusions.

First, the empirical results definitively confirm that single-layer defenses are inadequate for securing modern AI systems. The study revealed that standard deep learning models exhibit "Data Hypersensitivity," where a data poisoning rate of merely 3% causes a non-linear collapse in model reliability. Furthermore, the reliance on isolated metrics—such as accuracy alone—was proven dangerous, as high-confidence evasion attacks successfully bypassed standard models without detection.

Second, the proposed RAM-AI model successfully identified high-risk areas by transforming qualitative threats into quantitative scores. By integrating Asset Criticality, Threat Likelihood, and Model Vulnerability into a unified calculation, the framework provided a granular view of security posture. The evaluation of technical controls in Chapter 5 demonstrated that while specific interventions like Adversarial Training can restore robustness (recovering accuracy from 12% to 78% under attack), they often introduce trade-offs, such as increased computational overhead and slightly reduced clean-data accuracy. Consequently, the "Defense-in-Depth" architecture proposed in Chapter 6—combining input sanitization, algorithmic hardening, and Human-in-the-Loop (HITL) governance—was validated as the only viable strategy to mitigate the full spectrum of Data, Algorithmic, and Systemic risks.

7.2. Implications for Theory and Practice

The contributions of this research extend significantly to both the academic understanding of AI safety and its industrial application.

Theoretical Implications:

This study advances the field of AI safety metrics by bridging the gap between traditional cybersecurity risk assessment and machine learning stochasticity. By proving that "robustness" and "fairness" can be quantified and integrated into a unified risk index, this research challenges the prevailing binary view of security (secure vs. insecure). It contributes to the emerging discipline of AI Safety Engineering by providing empirical evidence of the "Accuracy-Robustness Trade-off" and defining the mathematical boundaries of perturbation tolerance.

Managerial Implications (A Playbook for CISOs):

For Chief Information Security Officers (CISOs), this dissertation provides a strategic playbook for navigating the transition from traditional IT security to AI security.

1) Quantifiable ROI: The RAM-AI model allows security leaders to translate abstract AI risks into business metrics, justifying the budget for computationally expensive defenses like adversarial training.
2) Operational Governance: The validated "Layered Architecture" (Input filtering leading to Robust Model leading to Human Oversight) offers a blueprint for compliance with emerging regulations such as the EU AI Act.
3) DevSecOps Integration: The findings support the shift to "MLSecOps," demonstrating that security gates (e.g., model signing and dependency scanning) must be automated within the CI/CD pipeline to prevent supply chain attacks.

7.3. Limitations

While this study offers a robust framework for AI risk assessment, specific limitations must be acknowledged regarding the scope and generalizability of the findings.

1) Data Modality Constraints: The experimental validation focused primarily on Computer Vision (image classification) and Tabular Data (financial fraud detection). While these represent two of the most critical deployment areas, the unique vulnerabilities associated with Natural Language Processing (NLP) or Audio processing—such as token manipulation or audio waveform perturbations—were not empirically tested in this iteration.
2) Computational Costs: As noted in Chapter 5, the implementation of robust adversarial training increased the model training time by 300% to 400%. This high computational cost may limit the applicability of the full defense framework in resource-constrained environments, such as edge computing or mobile devices, where latency and power consumption are critical bottlenecks.
3) Static Assessment: The RAM-AI model currently operates as a snapshot assessment. While valuable, it may not fully capture "Concept Drift" in real-time without frequent re-calibration, which can be operationally intensive.

7.4. Future Research Directions

Building upon the foundations laid by this dissertation, future research should expand into the following areas to address the evolving threat landscape:

1) Automated Defense Agents: Future work should explore the use of Reinforcement Learning (RL) to create autonomous defense agents. These agents could dynamically adjust defense parameters (e.g., the privacy budget in Differential Privacy or the filtering threshold in Input Sanitization) in real-time response to detected attack patterns, moving beyond static configurations.
2) Security for Large Language Models (LLMs): Given the explosive growth of Generative AI, there is an urgent need to adapt the RAM-AI framework for LLMs. Future research must investigate specific threats such as Prompt Injection, Jailbreaking, and Hallucination Induction. Developing quantitative metrics to measure the "semantic robustness" of LLMs—rather than just pixel-level robustness—will be critical for the safe deployment of next-generation AI agents.

References

1. Sarkar, S., Sunheriya, N., Giri, J., Al-Qawasmi, K., & Chadge, R. (2025). A Comprehensive Quantitative Model for Ethical AI Risk Assessment: EU Act on Artificial Intelligence. In Artificial Intelligence in the Digital Era: Economic, Legislative and Media Perspectives (pp. 145-165). Cham: Springer Nature Switzerland.
2. Grosse, K., Bieringer, L., Besold, T. R., Biggio, B., & Krombholz, K. (2023). Machine learning security in industry: A quantitative survey. IEEE Transactions on Information Forensics and Security, 18, 1749-1762.
3. Eckhart, M., Brenner, B., Ekelhart, A., & Weippl, E. (2019). Quantitative security risk assessment for industrial control systems: Research opportunities and challenges.
4. Hellas, M. S., Chaib, R., & Verzea, I. (2020). Artificial intelligence treating the problem of uncertainty in quantitative risk analysis (QRA). journal of engineering, design and technology, 18(1), 40-54.
5. Piorkowski, D., Hind, M., & Richards, J. (2025). Quantitative ai risk assessments: Opportunities and challenges. Seton Hall J. Legis. & Pub. Pol'y, 49, 644.
6. Murray, M., Barrett, S., Papadatos, H., Quarks, O., Smith, M., Boria, A. T., ... & Campos, S. (2025). A Methodology for Quantitative AI Risk Modeling. arXiv preprint arXiv:2512.08844.
7. Barrett, S., Murray, M., Quarks, O., Smith, M., Kryś, J., Campos, S., ... & Papadatos, H. (2025). Toward Quantitative Modeling of Cybersecurity Risks Due to AI Misuse. arXiv preprint arXiv:2512.08864.
8. Siddiqui, S. A., Thapa, C., Wang, D., Holland, R., Shao, W., Camtepe, S., ... & Shah, R. (2025). TELSAFE: Security Gap Quantitative Risk Assessment Framework. arXiv preprint arXiv:2507.06497.
9. Paracha, A., & Arshad, J. (2025). A bibliometric study toward quantitative research assessment of security of machine learning. Information Discovery and Delivery, 53(4), 481-498.
10. Grosse, K., & Alahi, A. (2024). A qualitative AI security risk assessment of autonomous vehicles. Transportation Research Part C: Emerging Technologies, 169, 104797.
11. Crotty, J., & Daniel, E. (2022). Cyber threat: its origins and consequence and the use of qualitative and quantitative methods in cyber risk assessment. Applied Computing and Informatics, (ahead-of-print).

12. Juric, M., Sandic, A., & Brcic, M. (2020, September). AI safety: state of the field through quantitative lens. In 2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO) (pp. 1254-1259). IEEE.

Disclaimer/Publisher's Note: The views, opinions, and data expressed in all publications are solely those of the individual author(s) and contributor(s) and do not necessarily reflect the views of PAP and/or the editor(s). PAP and/or the editor(s) disclaim any responsibility for any injury to individuals or damage to property arising from the ideas, methods, instructions, or products mentioned in the content.