ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

UNIVERSITY OF PIRAEUS

Thesis Defence – Department of Informatics – May 2025

Integrating Security by Design into Artificial Intelligence Systems

Student's name: GEORGIOS ZAIMIS

Supervisor: DESPINA POLEMI

About me


verizon
yahoo!
NFL
RYOT
forwARdgame
redlizard
STUDIOZ
VR/AR
ASSOCIATION
CROWD
POLICY
Athabasca
University
FACULTY OF BUSINESS
UNIVERSITY OF
PATRAS
ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
POS4work
POINT OF SYNERGY
ByeBye
Plastic
Foundation

Motivation & Objectives

Understanding the need for security in AI systems

Motivation & Objectives

Motivation and problem statement

Modern AI systems face unique security challenges like adversarial attacks and data breaches.

Traditional security methods often fail to address these AI-specific vulnerabilities.

Many AI applications lack security measures until after deployment, leaving them exposed.

Motivation & Objectives

Aim and purpose

To embed security from the start of AI system development.

Develop a systematic “Security by Design” approach for AI.

Improve resilience against threats without sacrificing performance.

Key objectives

1. Identify AI-specific security vulnerabilities and challenges not handled by conventional security practices.
2. Examine and adapt Security-by-Design principles to effectively address AI systems' needs.
3. Develop a framework to integrate security measures throughout the AI development lifecycle (design, implementation, testing, deployment, maintenance).
4. Apply the proposed framework in a case study on an existing AI application to demonstrate its feasibility.
5. Evaluate the impact of integrating security on the AI system's security posture and performance.

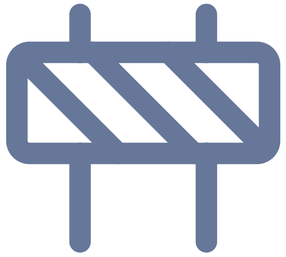
Theoretical Background

Exploring AI security challenges and principles

Theoretical Background

AI security challenges

Complex &
Dynamic Models



Adversarial Attacks

Bias & Ethical Issues

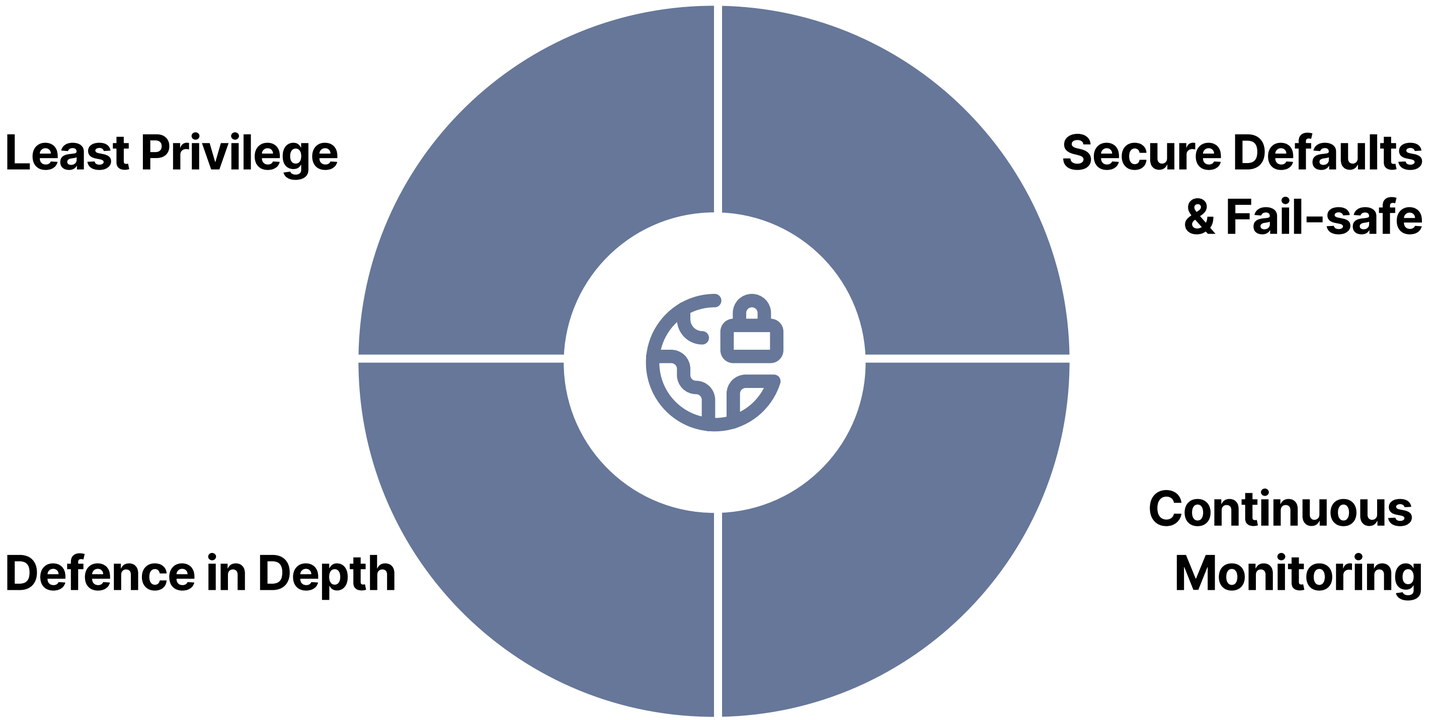
Data Privacy &
Integrity

Why Traditional Security Falls Short?

Theoretical Background

Security by Design Principles in AI

Security by Design (SbD): Philosophy of building systems secure from the ground up, rather than reacting to threats later. For AI, this means incorporating security considerations into every stage of model and system development. Key SbD principles include:



Methodology & Analysis

Developing and validating the Security-by-Design framework

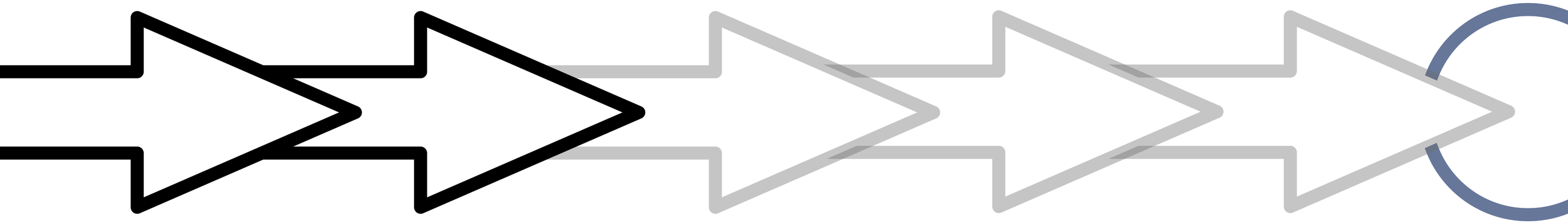
Methodology & Analysis

Research Questions

1. **AI Security Challenges:** What are the unique security challenges inherent in AI systems, and how do they differ from those in traditional software applications?
2. **Adapting SbD for AI:** How can Security-by-Design principles be effectively adapted to meet the specific security needs of AI systems?
3. **Integration Framework:** What kind of framework can be developed to systematically integrate SbD into the AI development process?
4. **Practical Implementation:** How can the proposed SbD framework be practically implemented in an existing AI application, and what challenges might be encountered during this integration?
5. **Impact on Security & Performance:** What is the impact of integrating Security-by-Design principles on the security and performance of AI systems (e.g. does security improve significantly, and what is the performance cost)?

Methodology & Analysis

Methodology



Research Design

Framework Development

Case Study Implementation

Evaluation Process

Tools & Environment

Methodology & Analysis

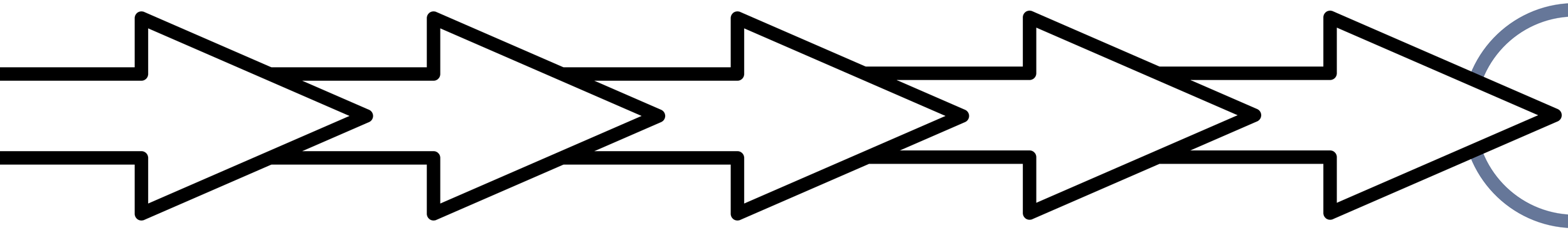
Framework Development

Researching on a complex codebase featuring data loaders, model definitions, training loops, evaluation scripts, and visualisation tools. Methodically created, initial steps to apply Security by Design principles:

Step 1: Architectural Review and Dependency Analysis
Step 2: Threat Modelling and Vulnerability Identification
Step 3: Security Requirements Derivation
Step 4: Tooling and Technique Selection
Step 5: Implementation Roadmap

Methodology & Analysis

Methodology



Research Design

Framework Development

Case Study Implementation

Evaluation Process

Tools & Environment

Data & Analysis

Data

- Data Handling
- Security Testing Procedures

Analysis

- Metrics Collected
  - Security Effectiveness
  - Performance Impact
- Analysis Methods

Key Findings

How the framework applies and performs in the AI stack

Key Findings

Security-by-Design Framework for AI

Framework Feasibility

Security Controls
Mapped to AI Lifecycle

Adaptation of Principles

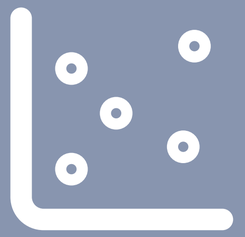
Key Findings

Case Study Implementation Results

Baseline vs Enhanced System

Security Improvements

Quantitative Outcomes

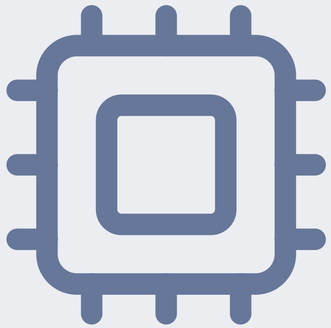


Non Functional Regression

Key Findings

Comparison with Unsecured Baseline System

Unsecure Baseline Characteristics



Security vs Speed

Case in Point

Key Findings

Comparison with Existing Approaches

In contrast, our Security-by-Design approach is embedded within the AI's workflow. According to our findings, this led to better outcomes:

Lower Latency

Broader Protection

Literature Benchmark

Conclusion & Contributions

We developed a comprehensive Security-by-Design framework for AI, adapting classic security principles to the specific context of AI systems. This framework is a novel blueprint for researchers and practitioners to build security-aware AI from the ground up.

Step 1: Define Security Requirements and Objectives
Step 2: Map the Existing Architecture and Data Flows
Step 3: Conduct Threat Modelling
Step 4: Secure Data Handling and Provenance
Step 5: Validate and Sanitise Inputs and Outputs
Step 6: Harden the Model and Surrounding Pipeline
Step 7: Implement Secure Coding and Dependency Management
Step 8: Perform Adversarial Robustness Testing
Step 9: Establish Access Control and Authentication Mechanisms
Step 10: Monitor, Detect, and Log Anomalous Activities
Step 11: Prepare Incident Response and Model Recovery Procedures
Step 12: Maintain and Update Security Over Time

Limitations

- **Scope of AI Technologies**
- **Known Threat Focus**
- **Experimental Setting**
- **Resource Constraints**
- **Generality and Customisation**

Future Work

- **Broader AI Domains**
- **Emerging Threats**
- **Real-World Deployments**
- **Performance Tuning**
- **User Trust & Policy Integration**

Thank You

Questions?