# Towards security of AI/ML

Isabel Praça

*School of Engineering of the Polytechnic of Porto*

*Portugal*

ENISA AI Cybersecurity Conference | Jun 07, 2023 | Brussels, Belgium

# Adversarial Machine Learning

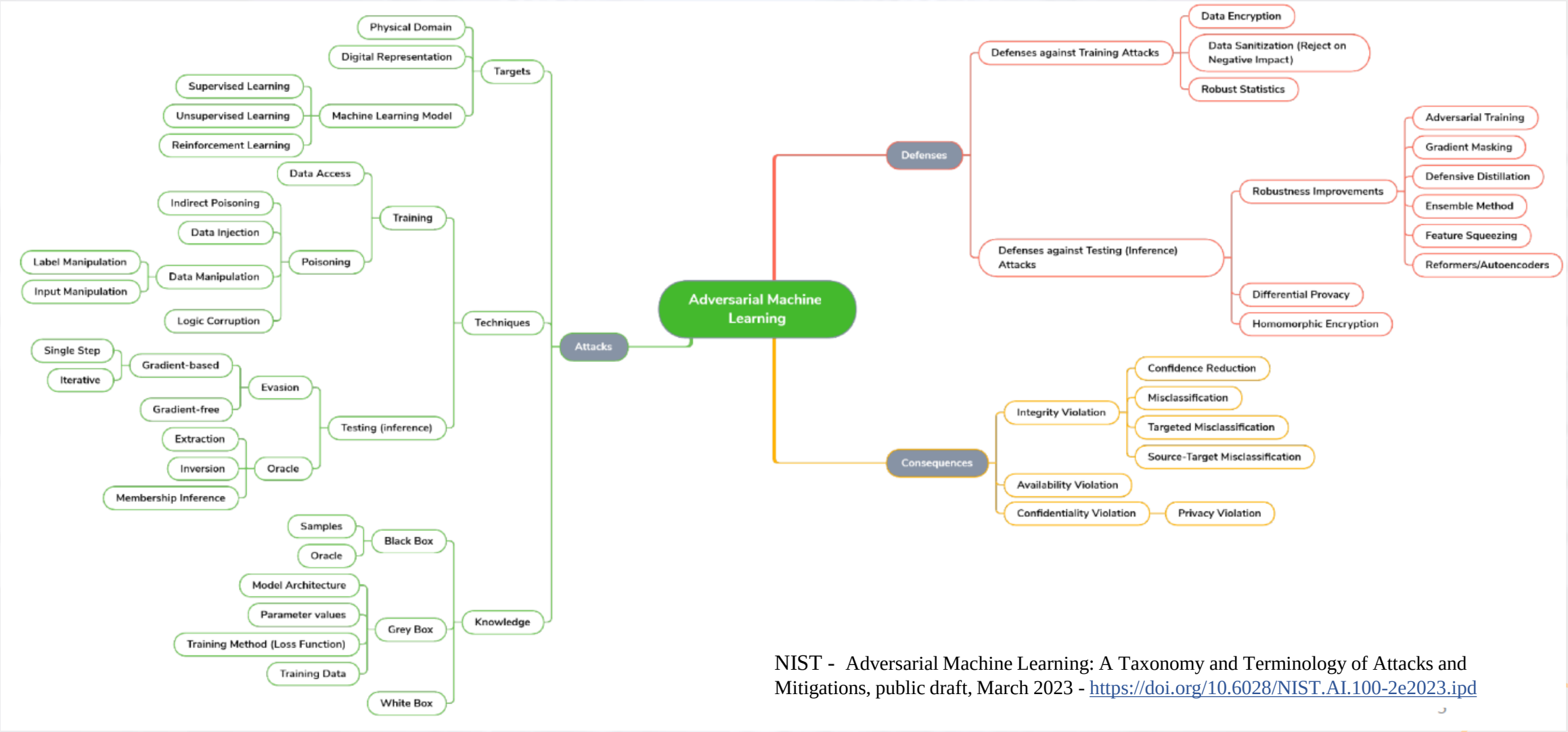
- AML is concerned with the design of ML algorithms that can resist security challenges, the study of the capabilities of attackers, and the understanding of attack consequences

*Attack (Target, Technique, Knowledge)*

*Defenses (Training Attacks, Inference Attacks)*

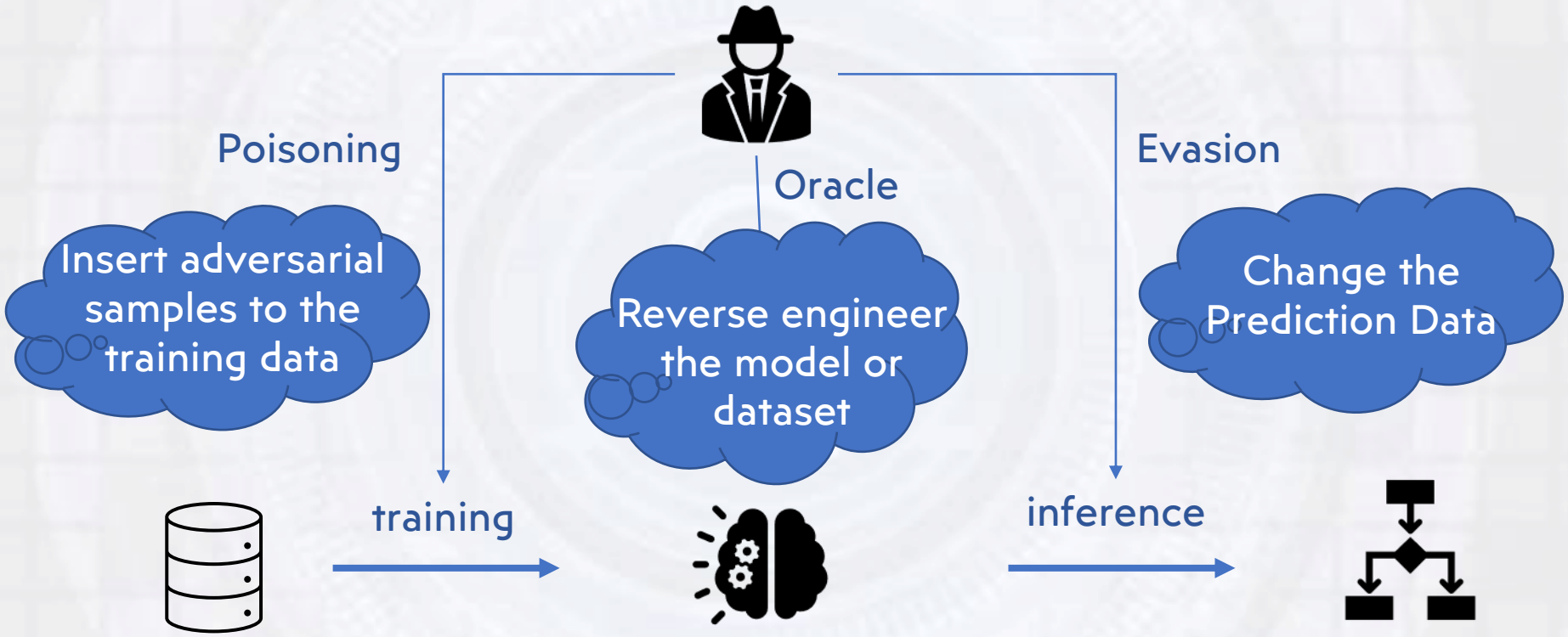
*Consequences (Confidentiality, Integrity, Availability)*

# Adversarial Machine Learning




NIST - Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations, public draft, March 2023 - https://doi.org/10.6028/NIST.AI.100-2e2023.ipd

# Attacks to AI/ML



# Desirable Properties

- Accuracy 
- Explainability 
- Fairness 
- Computational Efficiency 
- Robustness 
- Trustworthiness 

# Measuring Security and Trust of AI

## Trust & Security

**1.Robustness** to Attacks-related KPIs:
1. Adversarial Success Rate
2. Attack Detection Rate
3. Recovery Time
4. Out-of-Distribution Detection Rate

**2.Transparency**-related KPIs:
1. Transparency Score
2. Bias Detection Rate
3. Explanation Length

**3.Explainability**-related KPIs:
1. Model Explainability Score
2. Explainability Accuracy
3. User Satisfaction with Explanations

**4.Compliance and Ethical** Considerations-related KPIs:
1. Compliance Adherence Rate
2. Bias Mitigation Effectiveness

## Trust

**1.Fairness**-related KPIs:
1. Bias Detection Rate
2. Demographic Parity
3. Equalized Odds

**2.Reliability**-related KPIs:
1. Prediction Confidence
2. Prediction Consistency

**3.Generalization**-related KPIs:
1. Generalization Accuracy
2. Domain Adaptation Performance

**4.User Feedback and Satisfaction**-related KPIs:
1. User Satisfaction Score: Feedback provided by users on their satisfaction with the AI model's outputs, explanations, or overall performance.

## Security

**1.Privacy Preservation**-related KPIs:
1. Data Anonymization Effectiveness
2. Differential Privacy
3. Leakage Rate

**2.Authentication and Authorization**-related KPIs:
1. Authentication Success Rate
2. Access Control Effectiveness
3. Encryption Strength

**3.Adversarial Detection and Response**-related KPIs:
1. False Positive Rate
2. Detection Time
3. Attack Mitigation Success Rate

**4.Model Integrity**-related KPIs:
1. Model Tampering Detection Rate
2. Model Update Integrity
3. Model Rollback Prevention Rate

**5.Secure Data Handling**-related KPIs:
1. Data Encryption Effectiveness
2. Data Breach Incidents
3. Data Access Audit Accuracy

**6.Resilience to Data Poisoning**-related KPIs:
1. Poisoned Data Detection Rate
2. Model Performance Degradation
3. Data Sanitization Effectiveness

**7.Continuous Monitoring and Updates**-related KPIs:
1. Vulnerability Patching Time
2. Security Audit Completion Time
3. Incident Response Time

# New Metrics?

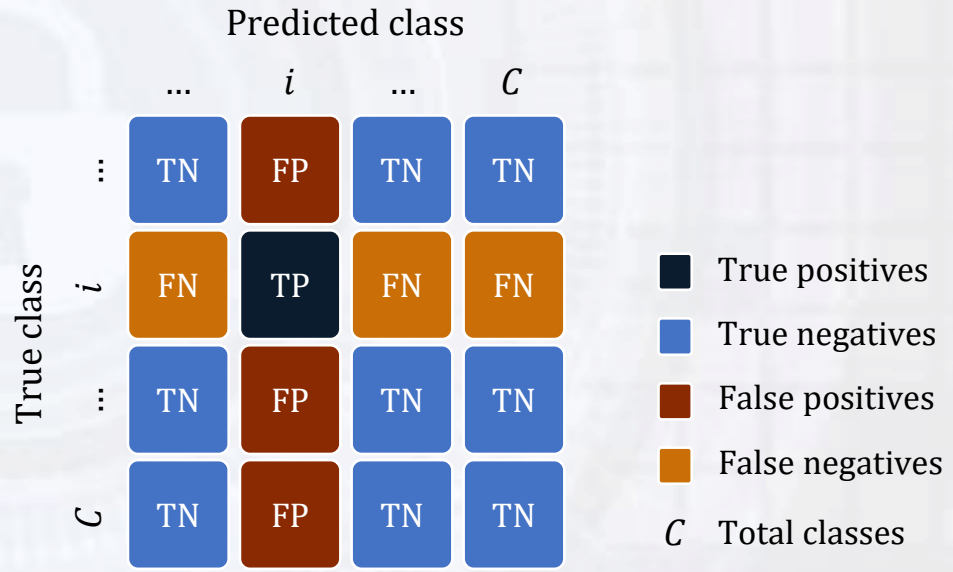
- Evaluating the robustness of ML models

Simulate an adversarial evasion attack vector targeting class $i$

Create perturbations in samples of other classes so they are misclassified as $i$

Analyze the increase of false positives of $i$, which denotes a lack of robustness

Improve the adversarial defense strategy and perform a new robustness analysis



# Metrics

## Balanced Accuracy

- Arithmetic mean of specificity and recall
- Defined as: $\frac{Specificity + Recall}{2}$

## F1-Score

- Harmonic mean of precision and recall
- Defined as: $\frac{2 * Precision * Recall}{Precision + Recall}$

## Specificity

- Proportion of samples of other classes that were correctly predicted as other classes
- Defined as: $\frac{TN}{TN + FP}$

## Recall

- Proportion of samples of class $i$ that were correctly predicted as class $i$
- Defined as: $\frac{TP}{TP + FN}$

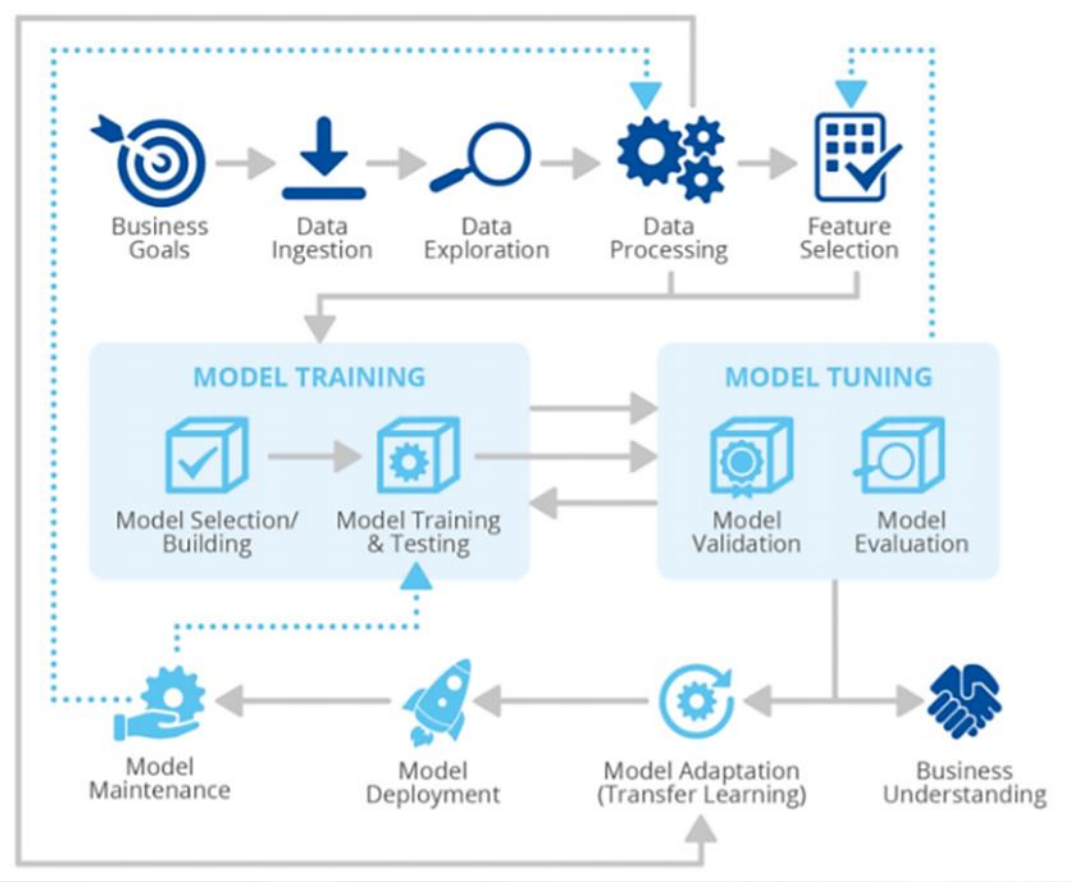
## Precision

- Proportion of samples predicted as class $i$ that were actually of class $i$
- Defined as: $\frac{TP}{TP + FP}$

## False Positive Rate

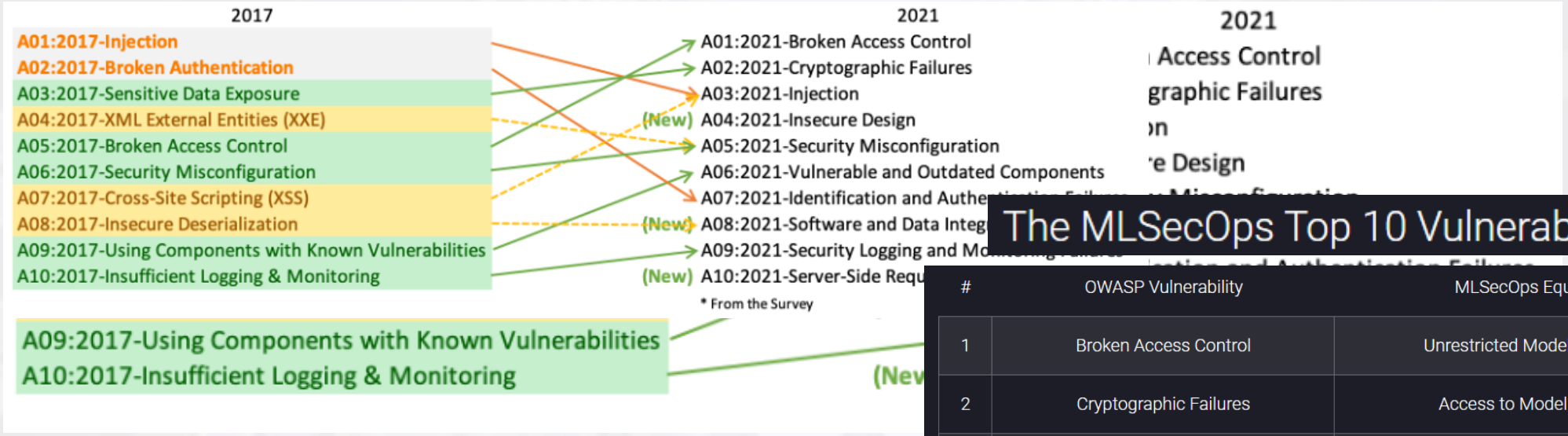
- Proportion of samples of other classes that were incorrectly predicted as class $i$
- Defined as: $\frac{FP}{FP + TN}$

# MLSecOps




- Integrates security practices into ML development and deployment
- Protects the privacy and security of training and testing data
- Safeguards deployed models and infrastructure from malicious attacks
- Implements secure coding practices and conducts threat modeling
- Performs security audits and establishes incident response for ML systems
- Ensures transparency and explainability to prevent unintended bias

# OWASP and MLSecOps




The MLSecOps Top 10

https://ethical.institute/security.html

## The MLSecOps Top 10 Vulnerabilities

| # | OWASP Vulnerability | MLSecOps Equivalent |
|---|---|---|
| 1 | Broken Access Control | Unrestricted Model Endpoints |
| 2 | Cryptographic Failures | Access to Model Artifacts |
| 3 | Injection | Artifact Exploit Injection |
| 4 | Insecure Design | Insecure ML Systems/Pipeline Design |
| 5 | Security Misconfigurations | Data & ML Infrastructure Misconfigurations |
| 6 | Vulnerable & Outdated Components | Supply Chain Vulnerabilities in ML Code |
| 7 | Identification & Auth Failures | IAM & RBAC Failures for ML Services |
| 8 | Software and Data Integrity Failures | ML Infra / ETL / CI / CD Integrity Failures |
| 9 | Logging and Monitoring Failures | Observability, Reproducibility & Lineage |
| 10 | Server-side Request Forgery | ML-Server Side Request Forgery |

# Threats to ML

- MITRE ATLAS™ (Adversarial Threat Landscape for Artificial-Intelligence Systems)

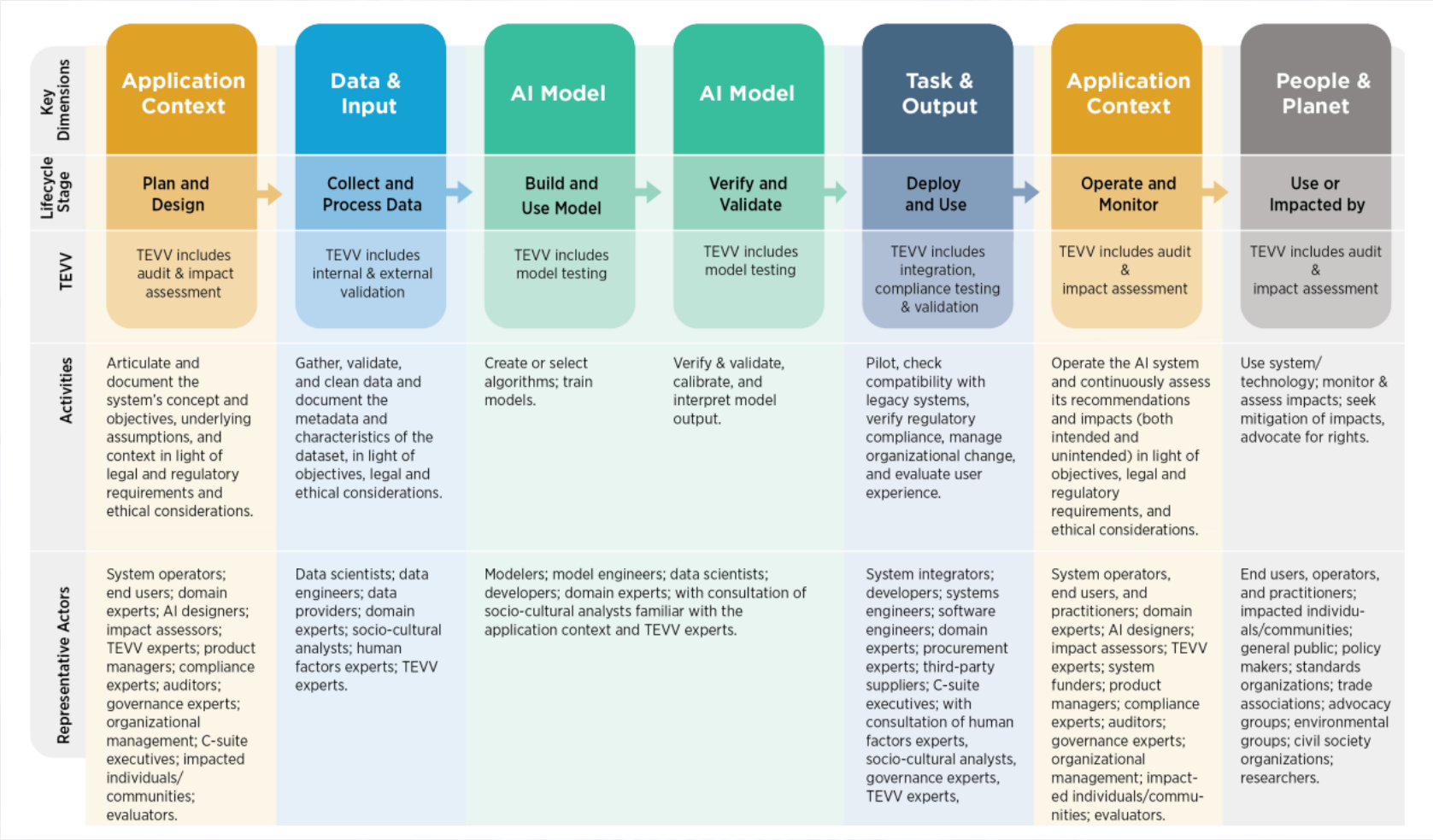
ATLAS™

The ATLAS Matrix below shows the progression of tactics used in attacks as columns from left to right, with ML techniques belonging to each tactic below. & indicates an adaptation from ATT&CK. Click on links to learn more about each item, or view ATLAS tactics and techniques using the links at the top navigation bar.

| Reconnaissance & | Resource Development & | Initial Access & | ML Model Access | Execution & | Persistence & | Defense Evasion & | Discovery & | Collection & | ML Attack Staging | Exfiltration & | Impact & |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 techniques | 7 techniques | 4 techniques | 4 techniques | 2 techniques | 2 techniques | 1 technique | 3 techniques | 3 techniques | 4 techniques | 2 techniques | 7 techniques |
| Search for Victim's Publicly Available Research Materials | Acquire Public ML Artifacts | ML Supply Chain Compromise | ML Model Inference API Access | User Execution & | Poison Training Data | Evade ML Model | Discover ML Model Ontology | ML Artifact Collection | Create Proxy ML Model | Exfiltration via ML Inference API | Evade ML Model |
| Search for Publicly Available Adversarial Vulnerability Analysis | Obtain Capabilities & | Valid Accounts & | ML-Enabled Product or Service | Command and Scripting Interpreter & | Backdoor ML Model | | Discover ML Model Family | Data from Information Repositories & | Backdoor ML Model | Exfiltration via Cyber Means | Denial of ML Service |
| Search Victim-Owned Websites | Develop Adversarial ML Attack Capabilities | Evade ML Model | Physical Environment Access | | | | Discover ML Artifacts | Data from Local System & | Verify Attack | | Spamming ML System with Chaff Data |
| Search Application Repositories | Acquire Infrastructure | Exploit Public-Facing Application & | Full ML Model Access | | | | | | Craft Adversarial Data | | Erode ML Model Integrity |
| Active Scanning & | Publish Poisoned Datasets | | | | | | | | | | Cost Harvesting |
| | Poison Training Data | | | | | | | | | | ML Intellectual Property Theft |
| | Establish Accounts & | | | | | | | | | | System Misuse for External Effect |

MITRE ATLAS™ and MITRE ATT&CK® are a trademark and registered trademark of The MITRE Corporation - https://atlas.mitre.org/

# AI Risk Management

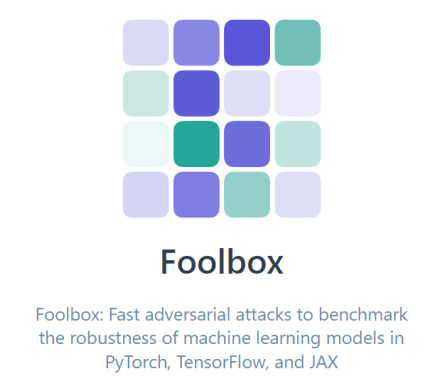


| Key Dimensions | Application Context | Data & Input | AI Model | AI Model | Task & Output | Application Context | People & Planet |
|---|---|---|---|---|---|---|---|
| Lifecycle Stage | Plan and Design | Collect and Process Data | Build and Use Model | Verify and Validate | Deploy and Use | Operate and Monitor | Use or Impacted by |
| TEVV | TEVV includes audit & impact assessment | TEVV includes internal & external validation | TEVV includes model testing | TEVV includes model testing | TEVV includes integration, compliance testing & validation | TEVV includes audit & impact assessment | TEVV includes audit & impact assessment |
| Activities | Articulate and document the system's concept and objectives, underlying assumptions, and context in light of legal and regulatory requirements and ethical considerations. | Gather, validate, and clean data and document the metadata and characteristics of the dataset, in light of objectives, legal and ethical considerations. | Create or select algorithms; train models. | Verify & validate, calibrate, and interpret model output. | Pilot, check compatibility with legacy systems, verify regulatory compliance, manage organizational change, and evaluate user experience. | Operate the AI system and continuously assess its recommendations and impacts (both intended and unintended) in light of objectives, legal and regulatory requirements, and ethical considerations. | Use system/ technology; monitor & assess impacts; seek mitigation of impacts, advocate for rights. |
| Representative Actors | System operators; end users; domain experts; AI designers; impact assessors; TEVV experts; product managers; compliance experts; auditors; governance experts; organizational management; C-suite executives; impacted individuals/ communities; evaluators. | Data scientists; data engineers; data providers; domain experts; socio-cultural analysts; human factors experts; TEVV experts. | Modelers; model engineers; data scientists; developers; domain experts; with consultation of socio-cultural analysts familiar with the application context and TEVV experts. | | System integrators; developers; systems engineers; software engineers; domain experts; procurement experts; third-party suppliers; C-suite executives; with consultation of human factors experts, socio-cultural analysts, governance experts, TEVV experts, | System operators, end users, and practitioners; domain experts; AI designers; impact assessors; TEVV experts; system funders; product managers; compliance experts; auditors; governance experts; organizational management; impact-ed individuals/commu-nities; evaluators. | End users, operators, and practitioners; impacted individu-als/communities; general public; policy makers; standards organizations; trade associations; advocacy groups; environmental groups; civil society organizations; researchers. |

NIST - Artificial Intelligence Risk Management Framework (AI RMF 1.0), January 2023 - https://doi.org/10.6028/NIST.AI.100-1

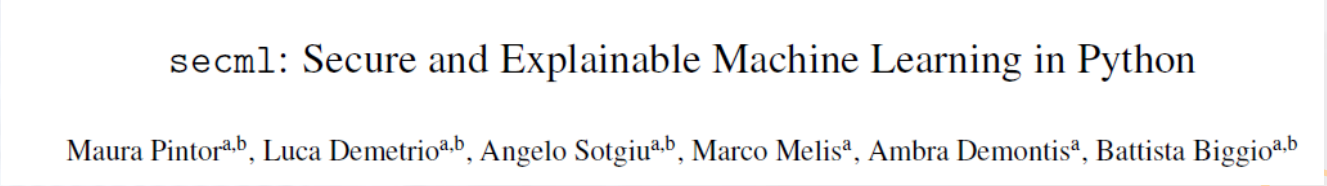
# Useful tools

- Foolbox - https://foolbox.jonasrauber.de/




- ART - https://github.com/Trusted-AI/adversarial-robustness-toolbox


- Secml – https://github.com/pralab/secml



# Way-ahead

# Thank You!

Isabel Praça

icp@isep.ipp.pt