Security in AI Applications

Safeguarding Data, Models, and Trust in the AI Ecosystem

Introduction

Security in AI ensures data confidentiality, model integrity, and reliable outputs. AI systems are both targets and tools for cyber attacks, expanding the attack surface.

Key Dimensions of AI Security

1. Data Security – protect training and input data
2. Model Security – prevent model theft/tampering
3. Pipeline Security – secure end-to-end ML lifecycle
4. Inference Security – protect deployed APIs and responses

Threat Landscape

- Data Poisoning: corrupting training data
- Model Inversion: reconstructing private data
- Adversarial Attacks: crafted inputs causing misclassification
- Model Extraction: logic theft via repeated queries
- Prompt Injection: manipulation of LLM behavior
- Supply Chain Attacks: malicious dependencies

Data Security

Encrypt data in transit and at rest; use differential privacy and role-based access.
Example (Differential Privacy):

```
from diffprivlib.models import LogisticRegression
model = LogisticRegression(epsilon=1.0)
model.fit(X_train, y_train)
```

Model Security

Protect weights with encryption or secure enclaves. Apply model watermarking and restrict access through authenticated APIs. Example: using AWS KMS for encryption.

Adversarial Defense Techniques

Approaches:
- Adversarial training
- Gradient masking
- Input preprocessing

Code Example:

```
from cleverhans.attacks import fgsm
adv_x = fgsm(model, X_test, eps=0.1)
```

Security in LLMs

Vulnerabilities: prompt injection, jailbreaks, context leakage.
Defenses: input sanitization, guardrails, output filters, retrieval isolation.

Secure ML Lifecycle

Data Collection → Model Training → Deployment → Inference
Each stage has risks; apply validation, encryption, authentication, and monitoring.

Compliance and Governance

Adopt frameworks like EU AI Act, NIST AI RMF, ISO/IEC 23894, and GDPR.
Maintain audit logs, explainability, and data protection policies.

Tools and Frameworks

- IBM Adversarial Robustness Toolbox (ART)
- TensorFlow Privacy
- Microsoft Presidio
- HuggingFace Guardrails
- Secure MLOps: MLflow + Vault + Kubernetes RBAC

Case Studies

1. Tesla Autopilot attack – lane misclassification
2. ChatGPT prompt injection examples
3. Healthcare AI model inversion leak

Best Practices Checklist

✔ Data encryption and anonymization
✔ Regular adversarial testing
✔ Secure APIs
✔ Model watermarking
✔ Access management
✔ Logging & anomaly detection
✔ Compliance audits

Future of AI Security

Emerging areas: AI red teaming, federated learning, quantum-safe encryption, AI-driven security operations (AI4SecOps).

Summary

Security in AI covers data, model, and deployment phases. Adopt DevSecOps practices and continuous monitoring.

Q&A; / Discussion

Prompt: If your model’s output could be manipulated, how would you detect and stop it?