What is AI Security?

What is AI Security?

AI Security encompasses the protection of artificial intelligence systems, their data, and the infrastructure that supports them from various threats and vulnerabilities.

Key Components

Model Security

Protecting AI models from adversarial attacks, data poisoning, and model extraction attempts.

Data Security

Ensuring the confidentiality, integrity, and availability of training and inference data.

Infrastructure Security

Securing the computational resources, APIs, and deployment environments for AI systems.

Common Threats

- **Adversarial Attacks**: Manipulating inputs to cause incorrect outputs
- **Data Poisoning**: Corrupting training data to compromise model behavior
- **Model Extraction**: Stealing model architecture and parameters
- **Prompt Injection**: Manipulating AI system behavior through crafted inputs

Best Practices

1. **Input Validation**: Validate and sanitize all inputs to AI systems
2. **Output Filtering**: Implement controls on AI system outputs
3. **Access Controls**: Restrict access to AI models and data
4. **Monitoring**: Continuously monitor AI system behavior and performance
5. **Regular Updates**: Keep AI systems and security measures current

Implementation Strategy

Start with a comprehensive risk assessment of your AI systems, then implement security controls based on identified threats and vulnerabilities.