CSA AI 系统安全指南

网络安全局 (CSA)

2024 年 10 月

中文翻译版

CSA AI 系统安全指南（Guidelines on Securing AI Systems）

概述

2024 年 10 月，新加坡网络安全局（Cyber Security Agency of Singapore, CSA）发布了《AI 系统安全指南》及配套实践手册，填补了 AI 安全领域的治理空白。这是新加坡首份专门针对 AI 系统安全的综合指南。

AI 系统全生命周期安全

1. 规划与设计阶段

- **威胁建模**：识别 AI 系统可能面临的安全威胁
- 安全需求分析和风险评估
- 安全架构设计原则
- 供应链安全考量

2. 开发阶段

- **数据安全**：确保训练数据的完整性和保密性
- **模型安全**：防止模型被篡改或窃取
- 安全编码实践
- 代码审查和安全测试

3. 部署阶段

- **安全测试**：部署前的全面安全评估
- 渗透测试和红队评估
- 安全配置和加固
- 访问控制和身份认证

4. 运维阶段

- **持续监控**：实时监测 AI 系统行为异常
- **事件响应**：建立 AI 安全事件应急机制
- 定期安全审计和评估
- 漏洞管理和补丁更新

AI 特有安全风险

对抗性攻击（Adversarial Attacks）

- 通过精心设计的输入欺骗 AI 系统
- 输入扰动导致错误输出
- 防御策略：对抗训练、输入验证、鲁棒性测试

数据投毒（Data Poisoning）

- 在训练数据中注入恶意数据
- 导致模型学习错误模式
- 防御策略：数据验证、异常检测、数据来源追溯

模型窃取（Model Extraction）

- 通过大量查询推断模型结构和参数
- 复制专有 AI 模型
- 防御策略：查询限制、输出扰动、访问控制

供应链安全（Supply Chain Security）

- 第三方模型和库的安全风险
- 预训练模型的后门风险
- 防御策略：供应链审计、模型验证、来源追溯

配套实践手册

CSA 同步发布了实践手册，为组织提供： - 具体的安全措施实施指南 - 行业案例和最佳实践 - 安全评估清单和模板 - 常见安全问题的解决方案

意义

该指南是新加坡 AI 治理体系在安全维度的重要补充。它与 AI 治理模型框架（侧重伦理治理）和 AI Verify（侧重合规测试）形成互补，共同构成了新加坡全方位的 AI 治理生态系统。