INTRODUCCIÓN A VULNERABILIDADES EN SISTEMAS DE IA

TICS00866: Auditoría y Defensa en Sistemas de Inteligencia Artificial

Documento: S5-1-02
Versión: 1.0
Fecha: Agosto 2025

¿QUÉ HACE VULNERABLE A UN SISTEMA DE IA?

Definición

Las **vulnerabilidades en sistemas de IA** son debilidades específicas que permiten a los atacantes comprometer la seguridad, integridad o funcionamiento de modelos de machine learning y sistemas inteligentes. A diferencia de las vulnerabilidades tradicionales de software, se enfocan en los aspectos únicos de los algoritmos de IA.

Diferencias con Vulnerabilidades Tradicionales

Aspecto	Vulnerabilidades Tradicionales	Vulnerabilidades de IA
Objeto de ataque	Código, redes, aplicaciones	Modelos, datos, algoritmos
Enfoque	Explotación de bugs	Manipulación de entrada/salida
Métricas	Disponibilidad, confidencialidad	Precisión, robustez, integridad
Riesgos	Acceso no autorizado, denegación de servicio	Clasificación incorrecta, extracción de datos
Técnicas	Buffer overflow, SQL injection	Adversarial attacks, model inversion

TIPOS DE VULNERABILIDADES EN IA

1. Vulnerabilidades de Entrada (Input Vulnerabilities)

- **Adversarial Attacks:** Manipulación de datos de entrada para engañar al modelo
- **Data Poisoning:** Inyección de datos maliciosos en el entrenamiento
- **Evasion Attacks:** Modificación de entrada para evadir detección

2. Vulnerabilidades de Modelo (Model Vulnerabilities)

- **Model Inversion:** Extracción de información del modelo entrenado
- **Membership Inference:** Determinación de si un dato fue usado en entrenamiento
- **Model Stealing:** Robo del modelo completo o parcial

3. Vulnerabilidades de Implementación (Implementation Vulnerabilities)

- **Inference Attacks:** Extracción de información durante la inferencia
- **Model Extraction:** Reconstrucción del modelo a través de consultas
- **Privacy Attacks:** Compromiso de privacidad de datos de entrenamiento

FRAMEWORKS DE SEGURIDAD EN IA

1. OWASP AI Exchange https://owaspai.org

- **Enfoque:** Seguridad en sistemas de IA
- **Categorías:** A01-A10 (vulnerabilidades específicas de IA)
- **Herramientas:** Checklist de evaluación

2. NIST AI Risk Management Framework https://www.nist.gov/itl/ai-risk-management-framework

- **Enfoque:** Gestión de riesgos en IA
- **Componentes:** Gobernanza, mapeo, medición, gestión
- **Proceso:** Iterativo y continuo

PROCESO DE PENTESTING EN SISTEMAS DE IA

Fase 1: Reconocimiento

1. **Identificar componentes:** Modelo, datos, API, infraestructura
2. **Mapear superficie de ataque:** Puntos de entrada, salidas, interfaces
3. **Recopilar información:** Documentación, configuración, logs

Fase 2: Análisis de Vulnerabilidades

1. **Identificar vectores de ataque:** Tipos de ataques aplicables
2. **Evaluar robustez:** Resistencia a ataques adversariales
3. **Analizar configuración:** Configuraciones de seguridad

Fase 3: Explotación

1. **Desarrollar exploits:** Crear ataques específicos
2. **Probar vulnerabilidades:** Ejecutar ataques en entorno controlado
3. **Documentar resultados:** Registrar hallazgos y evidencia

Fase 4: Post-Explotación

1. **Evaluar impacto:** Medir consecuencias de los ataques
2. **Desarrollar mitigaciones:** Proponer soluciones específicas
3. **Reportar hallazgos:** Documentar vulnerabilidades y recomendaciones

HERRAMIENTAS Y TÉCNICAS

Testing de Sistemas

- **Fuzzing:** Generación de entradas maliciosas
- **Penetration Testing:** Ataques simulados
- **Security Scanning:** Detección automática de vulnerabilidades

MEJORES PRÁCTICAS

Antes del Pentesting

- **Documentar todo:** Proceso, hallazgos, recomendaciones
- **Involucrar stakeholders:** Desarrolladores, operaciones, seguridad
- **Establecer línea base:** Estado actual del sistema

Durante el Pentesting

- **Ser sistemático:** Seguir metodología establecida
- **Ser ético:** No realizar ataques sin autorización
- **Ser documentado:** Registrar todo el proceso

Después del Pentesting

- **Seguimiento:** Verificar implementación de recomendaciones
- **Mejora continua:** Actualizar metodologías
- **Compartir aprendizajes:** Contribuir a la comunidad

CONCLUSIÓN

Las vulnerabilidades en sistemas de IA requieren un enfoque especializado que combine conocimientos de machine learning con técnicas de seguridad informática. El pentesting de sistemas de IA es esencial para garantizar la robustez y seguridad de estos sistemas críticos.

Puntos Clave:

- Las vulnerabilidades de IA son diferentes a las tradicionales
- Se enfocan en manipulación de entrada/salida y extracción de información
- Existen frameworks específicos para evaluación de seguridad
- El proceso debe ser sistemático y documentado
- Los hallazgos deben ser específicos y accionables

Referencias:

- OWASP AI Exchange: https://owasp.org/www-project-ai-exchange/
- NIST AI Risk Management Framework: https://www.nist.gov/ai-risk-management-framework
- Adversarial ML Taxonomy: https://arxiv.org/abs/1801.03984

Preparado por: Prof. Romina Torres

Fecha de preparación: Agosto 2025
Versión: 1.0