

From Black Box to Trustworthy AI: A Secure Framework for Explainable Cybersecurity Decision-Making

Presentation by **Adolfo Mora Córdova**

Noviembre 6, 2025

Introducción

La IA en ciberseguridad funciona como caja negra: es difícil validar el porque una alerta es disparada o como se deciden las acciones.

Lo cual reduce la confianza y complica las auditorias ya las mejoras del modelo.

Se requiere transparencia, seguridad y fiabilidad en entornos sensibles (IDS, SOC, etc.)

PROPOSED FRAMEWORK

Secure-XAI

1 Explicabilidad

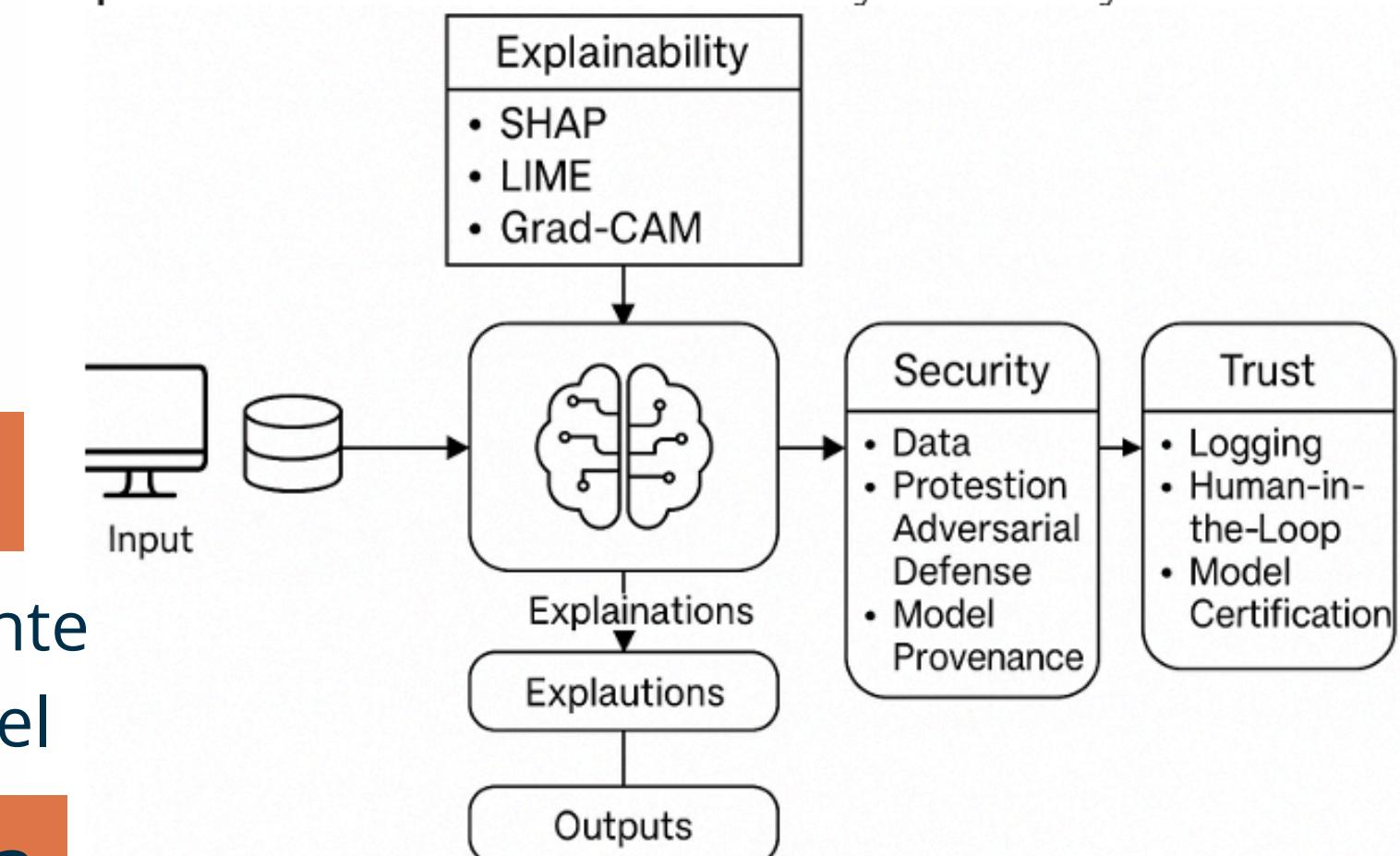
SHAP, LIME, Grad-CAM

2 Seguridad del sistema IA

Protección de datos, defensa ante adversarios, proveniencia/model

3 Mecanismos de confianza

Logging exhaustivo, Human-in-the-Loop, y certificación del modelo



Explicabilidad (XAI)

SHAP

Ofrece evaluaciones tanto globales como locales de la importancia de las características, lo que permite a los analistas de seguridad comprender la contribución de diversas características

LIME

Proporciona explicaciones localizadas al simular el comportamiento de un modelo complejo con un modelo más directo e interpretable, adaptado para individuales

Grad-CAM

Genera explicaciones visuales mediante la producción de mapas de calor que indican las regiones críticas dentro de los datos de entrada que influyeron en el proceso de toma de decisiones de la IA.

Seguridad del sistema de IA

Protección de datos

Cifrado de datasets y anonimización para reducir riesgos de filtración/privacidad.

Resiliencia ante adversarios

Entrenamiento adversarial + higienización de entradas (input sanitization).

Proveniencia del modelo

Trazabilidad detallada de datasets, versiones y procesos de entrenamiento para auditoría y detección de sesgos/vulnerabilidades.

Seguridad del sistema de IA

Logging

Exhaustivo de entradas,
salidas y racionales →
auditabilidad

Human-in-the-Loop

En decisiones críticas
para reducir riesgo de
sobre-dependencia en IA.

Certificación del modelo

Pruebas contra
benchmarks de
seguridad, explicabilidad
y fiabilidad para elevar la
credibilidad ante
stakeholders.

Casos de uso

Malware

Phishing

SOCs aumentados con IA

Conclusión

Secure-XAI mejora transparencia, confiabilidad y control; dota a analistas de medios para entender, validar y gobernar decisiones de IA en ciberseguridad.

Trabajo futuro propuesto

- Prototipo funcional
- Evaluación en Ambiente operativos reales

Thank You So Much!

Presentation by **Adolfo Mora Córdova**