"IA PARA LA CIBERSEGURIDAD" Y "CIBERSEGURIDAD PARA LA IA"

MOTIVACIÓN

El uso de sistemas basados en IA abre nuevas superficies de ataque contra los algoritmos de IA y aprendizaje computacional. Los componentes basados en IA suelen funcionar como cajas negras, convirtiéndose en objetivos ideales para atacantes que atentan contra la robustez de estos sistemas mediante el desarrollo de ataques de IA Adversarial. Estos ataques violan la integridad de los modelos mediante el uso de entradas maliciosas.

DESCRIPCIÓN

AIAS tiene como objetivo diseñar y desarrollar una plataforma de seguridad innovadora para proteger los sistemas de IA mediante soluciones de IA adversarial, técnicas de engaño y modelos de IA explicable, mejorando la resiliencia frente a los ciberataques.

ACCIONES

El principal objetivo de AIAS es realizar una investigación exhaustiva sobre la IA Adversarial de cara a diseñar y desarrollar una innovadora plataforma de seguridad basada en IA que proteja la robustez de los sistemas basados en IA. Para ello, AIAS se apoya en métodos de defensa contra IA Adversarial, mecanismos de engaño y soluciones de IA eXplicable (XAI).

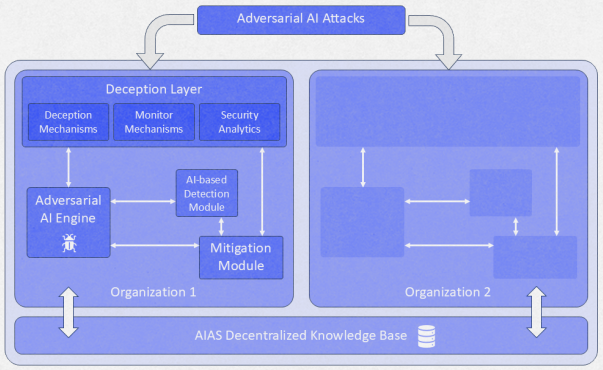
DICIEMBRE 2024

AIAS

PLATAFORMA DE CIBERSEGURIDAD ASISTIDA POR IA

HORIZON-MSCA-2022-SE-01-01;
HORIZON.1.2 - Marie Skłodowska-Curie Actions (MSCA)

PÁGINA WEB: https://www.aias-project.eu/
INICIO DEL PROYECTO: 1 de enero de 2024
DURACIÓN: 48 meses
ACUERDO: 101131292
FONDOS EUROPEOS: EUR 1 564 000
COORDINACIÓN: University of Piraeus Research Center (Grecia)



IMPORTANCIA DE LA IA EN AIAS

Motor de IA Adversarial
Generación de escenarios de ataque adaptados a las características de las organizaciones.

Redes Generativas Adversariales (GANs)
Detección de ataques de IA Adversarial.

XAI
Motor de recomendaciones para la mitigación de ataques y una mejor comprensión de las decisiones tomadas por la IA.

OBJETIVOS

PROTECCIÓN HOLÍSTICA

Conceptualizar y desarrollar una arquitectura donde servicios potenciados por IA, mecanismos de engaño y técnicas de mitigación trabajen juntos para la protección holística de las organizaciones contra los ciberataques y la IA Adversarial.

ESCENARIOS DE ATAQUE

Diseñar y desarrollar un motor de IA Adversarial novel para la creación de escenarios de ataque adaptados a las características hardware y software de las organizaciones objetivo.

MÉTODOS DE ENGAÑO INTELIGENTES

Implementación de métodos de engaño inteligentes basados en honeypots de alta interacción, gemelos digitales y personas virtuales.

PROTECCIÓN BASADA EN IA

Diseñar, desarrollar y evaluar métodos basados en IA para la detección y mitigación de ciberataques, así como implementar métodos de recopilación fusión de datos.

MOTOR DE RECOMENDACIONES BASADO EN IA EXPLICABLE

Desarrollar y verificar un motor de recomendaciones basado en XAI que facilite las decisiones proactivas de intervención humana con el objetivo de mitigar de forma exhaustiva los ataques de IA Adversarial.

ESCENARIOS DE USO REALES

Validar el funcionamiento, efectividad y eficiencia de AIAS en escenarios del mundo real.

METODOLOGÍA

Fase 1: Identificación de los principales componentes y requisitos del sistema.

- Identificación de los requisitos funcionales de seguridad y privacidad.
- Revisión exhaustiva del estado del arte en áreas claves, mecanismos de engaño y métodos de detección y mitigación potenciados por IA.
- Especificación de las herramientas y aplicaciones necesarias para la implementación de cada módulo de AIAS.

Fase 2: Implementación y validación de los componentes principales de la plataforma. Cada módulo se diseñará siguiendo técnicas reconocidas, implementando métodos y directrices proporcionadas por la UE, y utilizando tecnologías de vanguardia.

Fase 3: Integración, estudio de la prueba de concepto y evaluación en escenarios del mundo real. El objetivo principal de esta fase es concluir la plataforma AIAS, asegurando que los módulos que la componen sean funcionales y operen de forma correcta. Una vez integrada la plataforma, todos los socios procederán con la evaluación en casos pilotos del mundo real. Esta última tarea podría traducirse en modificaciones de la plataforma atendiendo a las críticas durante las pruebas.