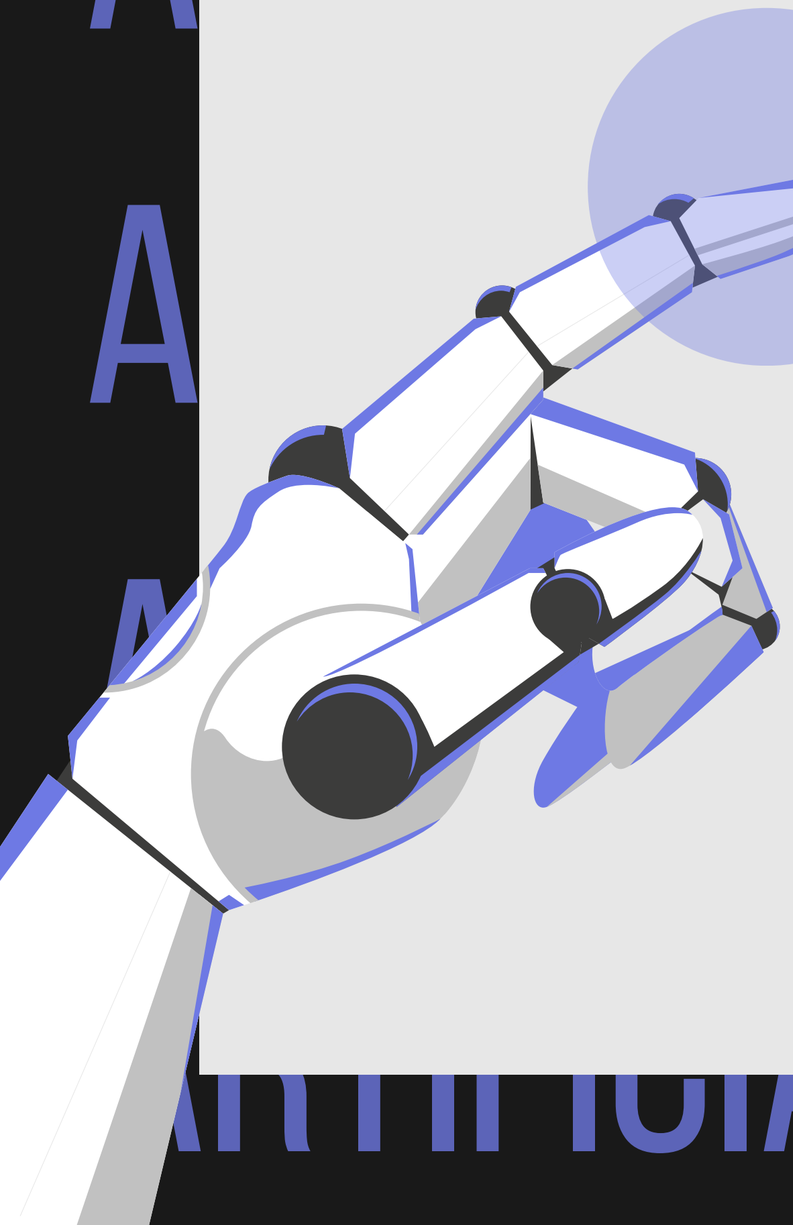


VULNERABILIDADES EN LA IA MODERNA

Basado en el artículo Understanding AI Vulnerabilities

-Martínez Lizarraga Abel Alejandro

INTRODUCCIÓN

ARTIFI
INTELL
(AI)

La IA moderna es poderosa, flexible y ampliamente utilizada.
Pero esta misma flexibilidad la hace vulnerable.
Nuevos tipos de ataques están apareciendo.
La seguridad en IA se vuelve cada vez mas crítica

<<<<

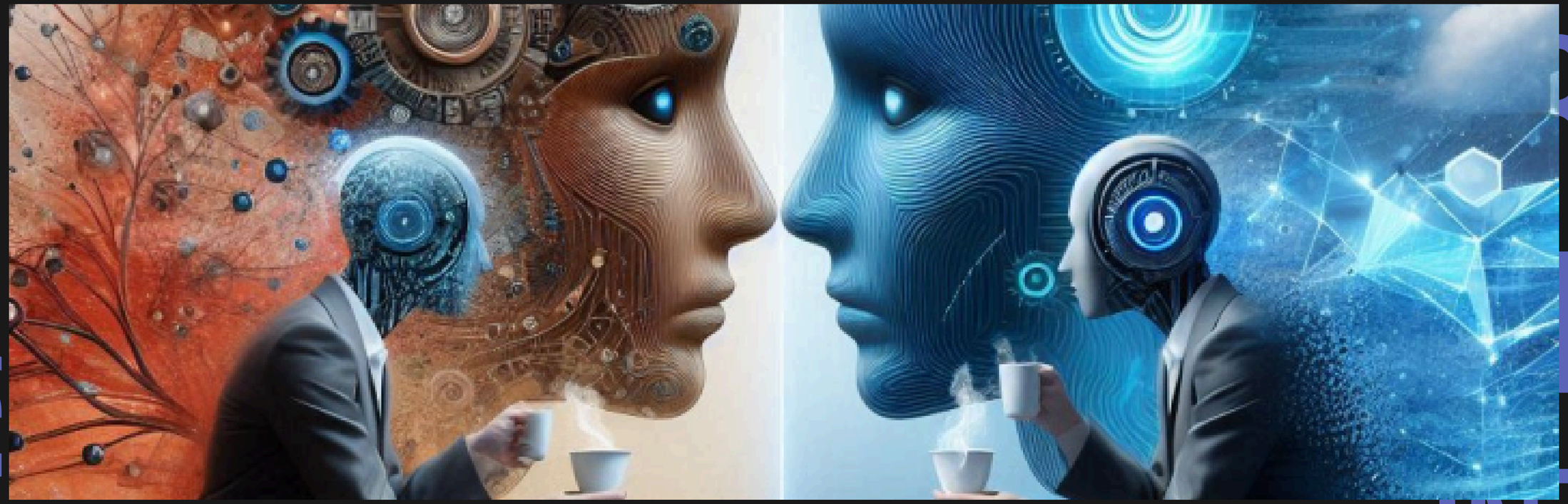
¿POR QUÉ LA IA ES VULNERABLE?

Mayor complejidad provoca mayor superficie de ataque

Modelos generativos pueden ser manipulados

Dependencia de datos masivos

Interacción mediante lenguaje natural



IA TRADICIONAL

- Tareas específicas
- Clasificación y predicción
- Riesgo limitado

IA GENERATIVA

- Produce contenido nuevo
- Interpreta lenguaje natural
- Más caminos para atacarla

ADVERSARIAL ATTACKS

ARTIFI
INTELL
(AI)

Cambios mínimos en el input provoca errores enormes en el output.
Pueden evadir filtros y engañar clasificadores.

- Usados para:

- Evadir detección
- Manipular decisiones
- Filtrar información

DATA POISONING

SI LOS DATOS ESTÁN CONTAMINADOS → la IA aprende mal

RIESGOS ←

- Respuestas sesgadas
- Fugas de información personal
- Comportamiento malicioso aprendido

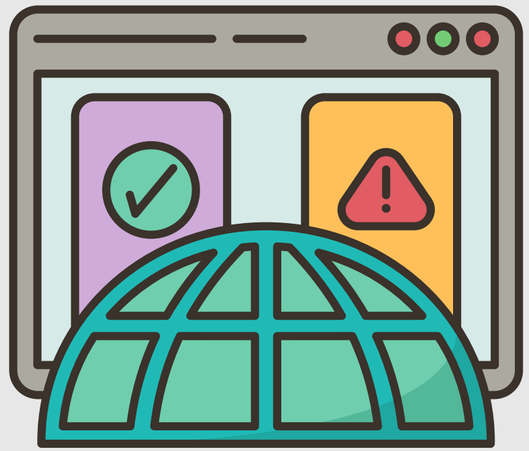
ARTIFICIAL INTELLIGENCE (AI)

PROMPT INJECTION

ARTI
INTEL
(AI)

Manipulación del modelo usando texto.
Permite:

- Saltar reglas
- Engañar políticas de seguridad
- Acceder a información privada
- Relacionado con ataques tipo jailbreak



JAILBREAKING

TÉCNICAS PARA ROMPER LAS RESTRICCIONES DE SEGURIDAD DEL MODELO

PERMITE

- CONTENIDO PELIGROSO
- INSTRUCCIONES ILEGALES
- DATOS SENSIBLES
- AFECTA INCLUSO A MODELOS AVANZADOS

ARTIFICIAL INTELLIGENCE (AI)

VULNERABILIDADES EN APIS

ARTI
INTEL
(AI)

Las empresas usan modelos generativos vía API.
Un error en el modelo puede provovar muchas empresas afectadas.

Riesgos:

- Exposición de clientes
- Ataques masivos
- Fugas de datos sensibles

SOLUCIONES

AI FIREWALL → SISTEMA QUE FILTRA INPUTS Y OUTPUTS

BLOQUEA →

- PROMPTS MALICIOSOS
- CÓDIGO DAÑINO
- DATOS SENSIBLES
- FUNCIONA COMO UN FIREWALL TRADICIONAL, PERO PARA IA

ARTIFICIAL INTELLIGENCE (AI)

VALIDACIÓN DE DATOS

REVISIÓN PREVIA AL ENTRENAMIENTO

DETECTA

- SESGOS
- INFORMACIÓN PERSONAL
- CONTENIDO MALICIOSO
- EVITA FALLOS ÉTICOS Y FILTRACIONES

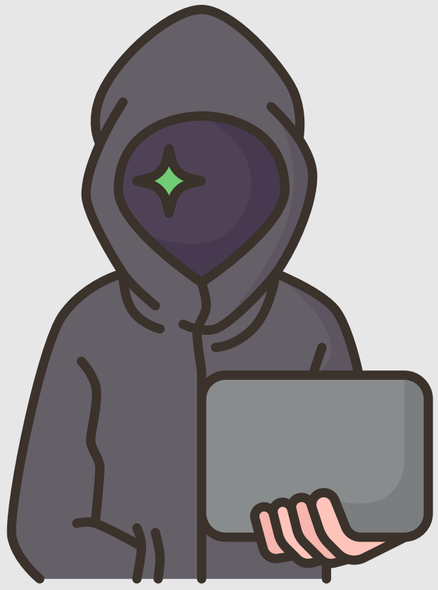
ARTIFICIAL INTELLIGENCE (AI)

ADVERSARIAL TESTING

ARTI
INTEL
(AI)

Pentesting aplicado a modelos de IA.
Simulan ataques reales.

- Permite identificar:
- Vulnerabilidades
- Comportamiento inesperado
- Fallos en outputs o instrucciones



CASO REAL

AR
IN
CA

VULNERABILIDADES EN GUARDRAILS

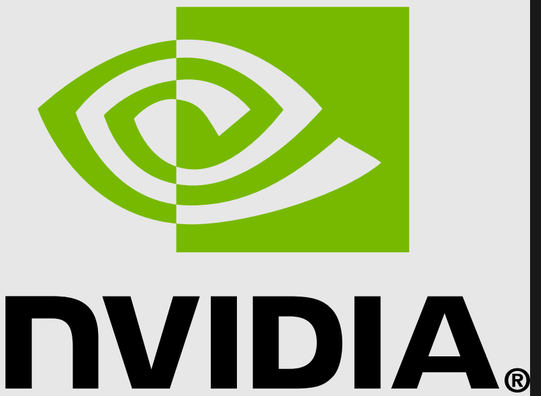
Investigaciones recientes demostraron que incluso los sistemas de protección (guardrails) pueden ser evadidos

NVIDIA FUE SOMETIDA A PRUEBAS ADVERSARIALES.

ARTI
INTEL
(AI)

Los atacantes lograron usar prompts diseñados para:

- Saltar filtros de contenido
- Extraer datos privados del modelo
- Obtener instrucciones peligrosas sin autorización



<<<<

AP
IN
CA

¿CÓMO SE VULNERARON?

- Combinando instrucciones ambiguas con mensajes encadenados
- Alterando inputs de forma que parecían seguros pero contenían lógica oculta
- Aprovechando debilidades en la interpretación del lenguaje natural

RIESGOS ÉTICOS

ARTIFICIAL

REPRODUCIR SESGOS

DISCRIMINAR

TOMAR DECISIONES INJUSTAS

<<<< REGULACIÓN Y RESPONSABILIDAD >>>>

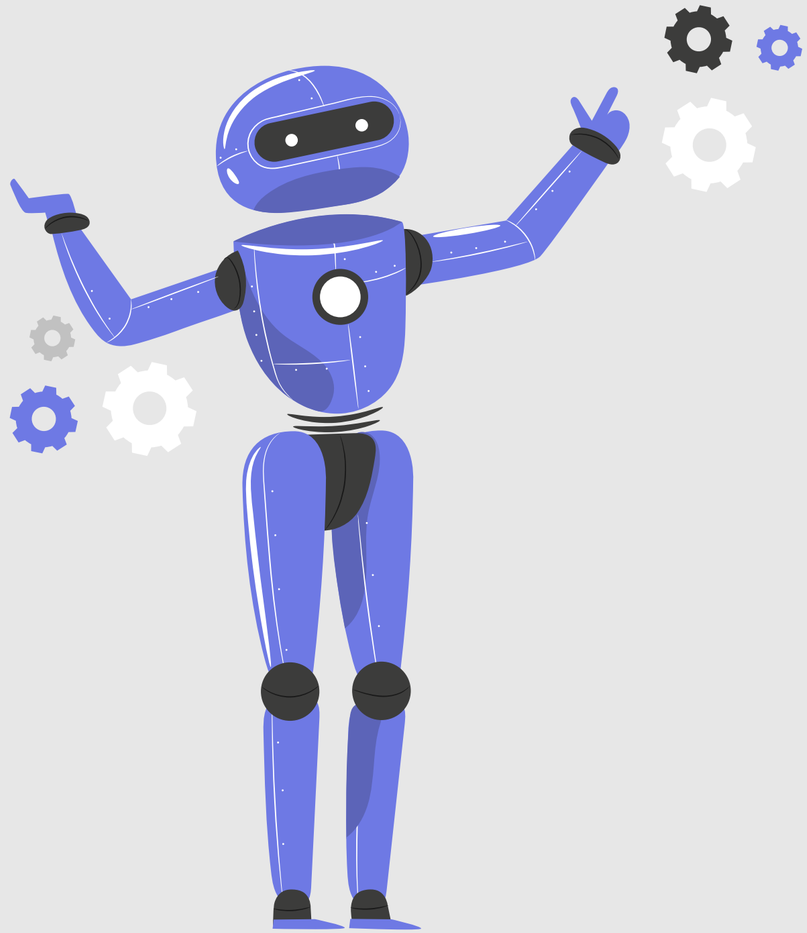
LAS EMPRESAS DEBEN IMPLEMENTAR

-GUARDRAILS

-IA SEGURA POR DISEÑO

-PROTECCIÓN DE DATOS

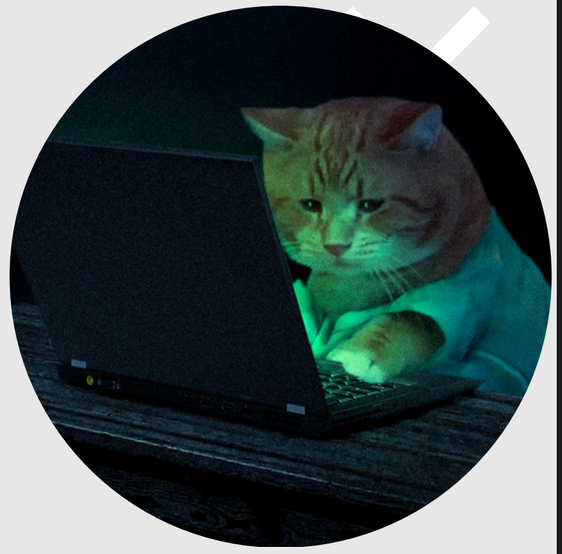
-REGULADORES EXIGEN TRANSPARENCIA Y SEGURIDAD



AR
IN
CA

UN JUEGO DE GATO Y RATÓN

- No existe seguridad absoluta
- Cada nueva defensa genera un nuevo ataque
- La seguridad debe adaptarse constantemente



AR
IN
CA

CONCLUSIÓN

La IA moderna es extremadamente poderosa.
Pero también es vulnerable a múltiples ataques.

La seguridad requiere:

- AI Firewalls
- Validación de datos
- Pruebas adversariales
- La protección será un proceso continuo

ARTICULO

- **www.harvardmagazine.com/2025/03/artificial-intelligence-vulnerabilities-harvard-yaron-singer**