

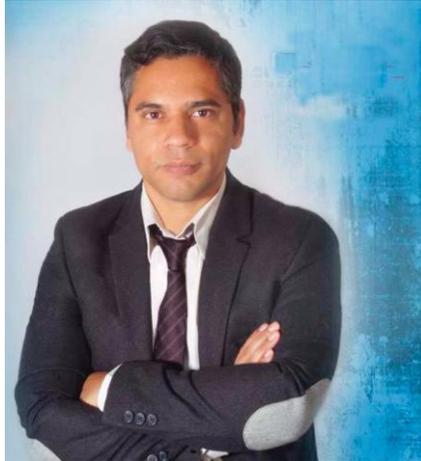
La ciberseguridad basada en herramientas de IA

30 de noviembre 2023
Adrián Ramírez

Junta de Andalucía



ABOUT ME



Adrián Ramírez



@adriandolbuck



Junta de Andalucía

Security Consultant
Data recovery
Computer forensics
Defense mitigation
Computer and networking Security
Social engineer
Graphic Design
Writer
Pentester
Organizer: SecAdmin
Docente del Máster de Seguridad Ofensiva
CISSP, CEH, OSCP



ÍNDICE DE LA UNIDAD

1. Qué es la IA
2. Cómo funciona la IA
3. Tipos de inteligencia artificial
4. Ejemplos de uso de la IA
5. Las IA más famosas
6. La ciberseguridad basada en herramientas con IA
 - IA del lado del mal
 - IA del lado del bien
7. Conclusiones

Qué es la IA

La inteligencia artificial (IA) es un campo de la informática que se enfoca en crear sistemas que puedan realizar **tareas que normalmente requieren inteligencia humana**, como el aprendizaje, el razonamiento y la percepción. Estos sistemas pueden percibir su entorno, razonar sobre el conocimiento, procesar la información derivada de los datos y tomar decisiones para lograr un objetivo dado.



Cómo funciona la IA

Las Inteligencias artificiales utilizan **algoritmos y modelos matemáticos** para procesar grandes cantidades de datos y tomar decisiones basadas en patrones y reglas establecidas a través del **aprendizaje automático**, que es la capacidad de una máquina para aprender de forma autónoma a partir de datos sin ser programada específicamente para hacerlo. De esta manera la IA puede mejorar su precisión y eficiencia con el tiempo.



Tipos de Inteligencia Artificial

- **Software:** asistentes virtuales, software de análisis de imágenes, motores de búsqueda o sistemas de reconocimiento de voz y rostro.
- **Inteligencia artificial integrada:** robots, drones, vehículos autónomos o el Internet de las Cosas.





Tabla periódica de

Suscríbete

Descárgala

+ Apps de IA

Versión 1.1

Apps gratuitas de Inteligencia Artificial

por @andreaoviedov www.bit.ly/tablaia





Ejemplo de uso de la IA



Las IAs más famosas

Durante la segunda mitad de 2022 y 2023 se ha producido una explosión en el campo de las inteligencias artificiales.

Algunos ejemplos de modelos que han visto la luz en los últimos dos años y que han acaparado una gran atención mediática son los siguientes:

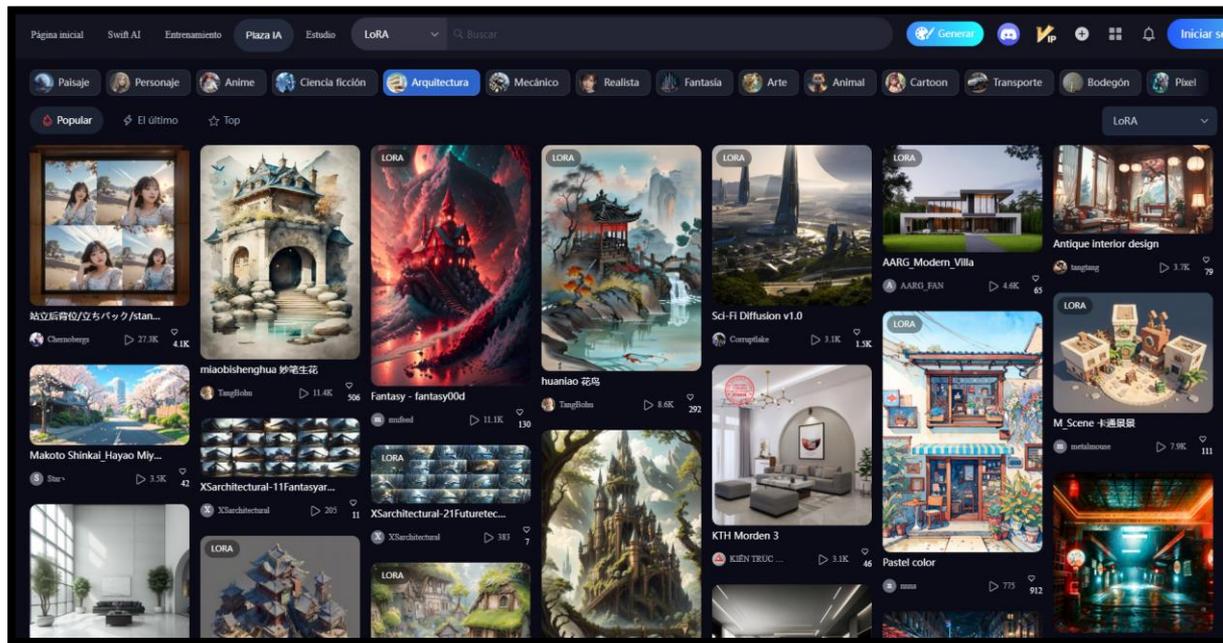
Procesamiento del Lenguaje Natural: [ChatGPT](#), [GPT-4](#), [BERT](#), [Bing Chat](#)

Generación de Imágenes: [DALL-E](#), [CLIP](#), [StyleGAN 3](#), [Stable Diffusion](#), [Midjourney](#)

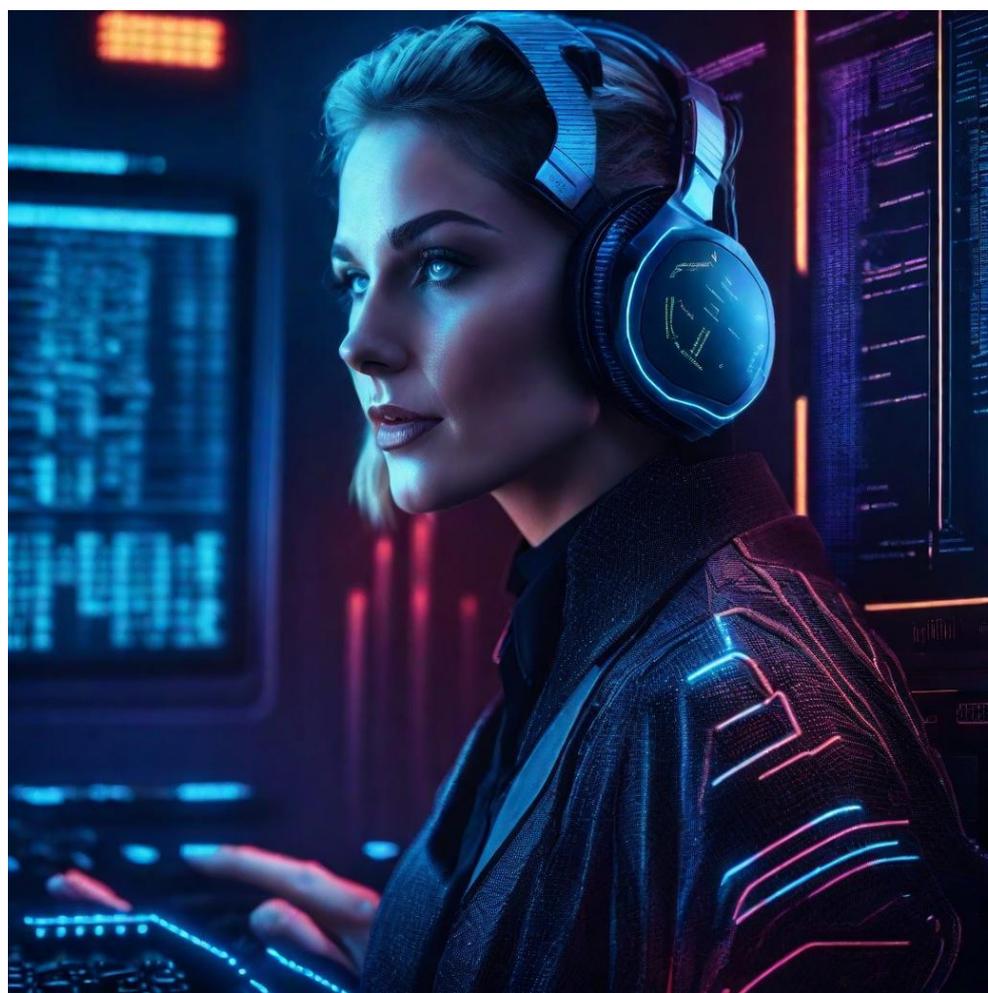
Transcripción de Audio: [DeepSpeech](#), [Wav2Vec 3](#), [QuartzNet](#), [Whisper](#)

Generación de Música: [MuseNet](#), [Amper Music](#), [AIVA](#)

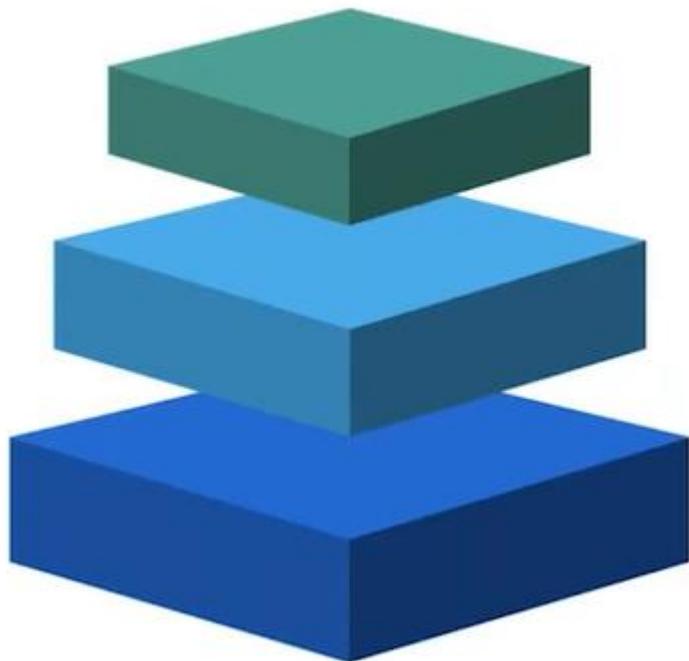
Generación de Código de Programación: [GPT-Codex](#), [GitHub Copilot](#), [Deep TabNine](#)



Ciberseguridad basada en herramientas con IA



Inteligencia Artificial y Ciberseguridad



Uso de la Inteligencia Artificial para mejorar las **técnicas de ataque** existentes



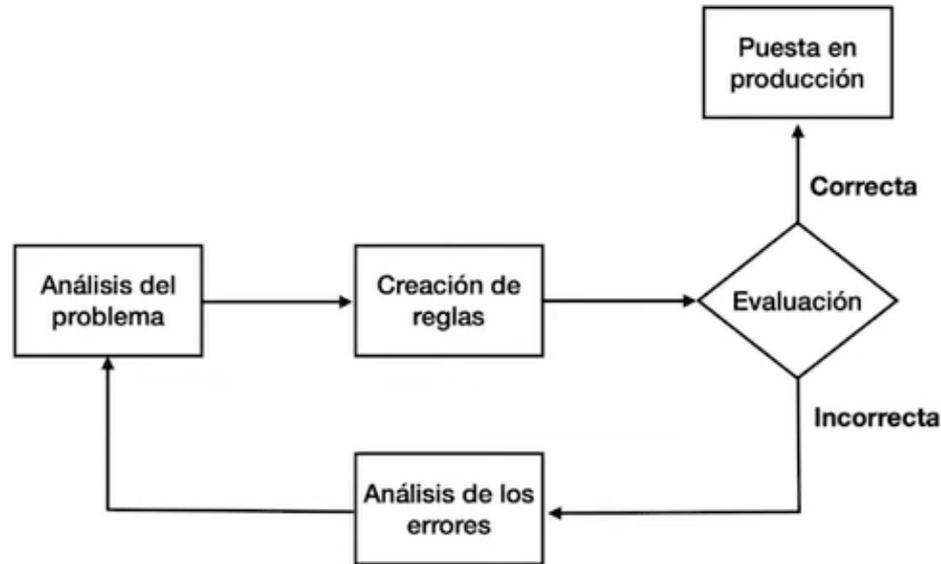
Uso de la Inteligencia Artificial para mejorar las **técnicas de defensa** existentes



Seguridad de los sistemas basados en Inteligencia Artificial

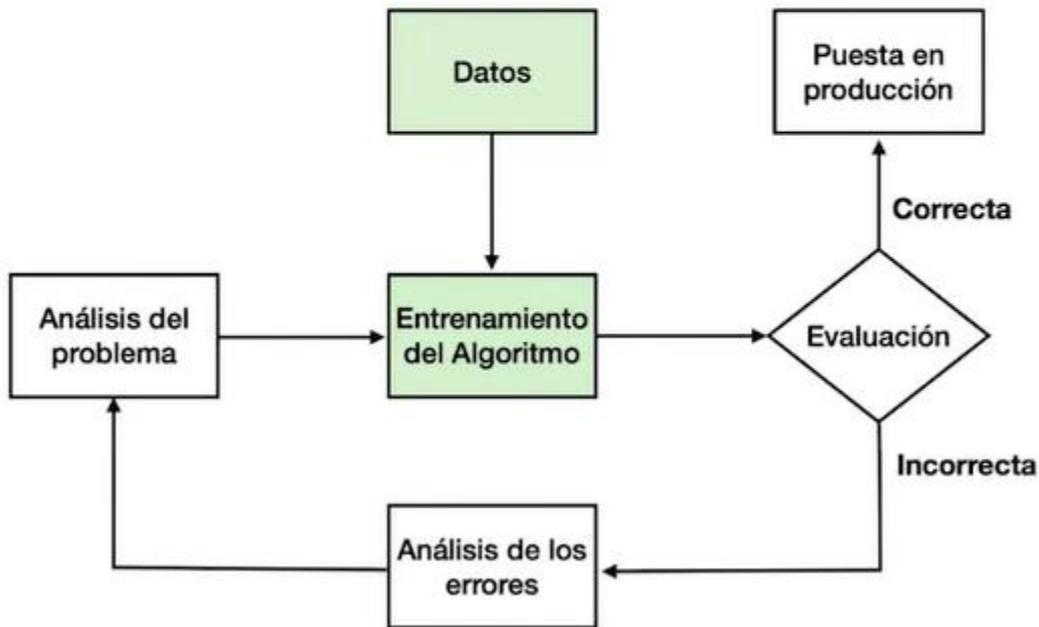
Inteligencia Artificial y seguridad defensiva

- ¿Cómo funcionan las herramientas de seguridad en la actualidad?



Inteligencia Artificial y seguridad defensiva

- Aplicación de Machine Learning a un proceso de seguridad



100 STARTUPS USING ARTIFICIAL INTELLIGENCE TO TRANSFORM INDUSTRIES

CONVERSATIONAL AI/ BOTS



VISION



AUTO



ROBOTICS



CYBERSECURITY



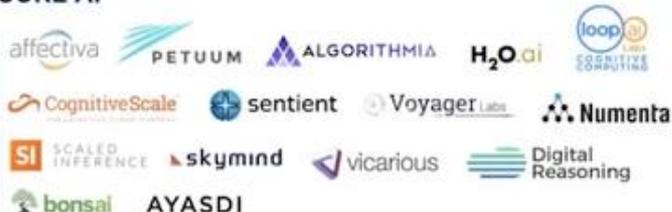
BUSINESS INTELLIGENCE & ANALYTICS



AD, SALES, CRM



CORE AI



HEALTHCARE



TEXT ANALYSIS/ GENERATION



IOT/IIOT



COMMERCE



FINTECH & INSURANCE



OTHER



THE INNOVATION SUMMIT

CBINSIGHTS

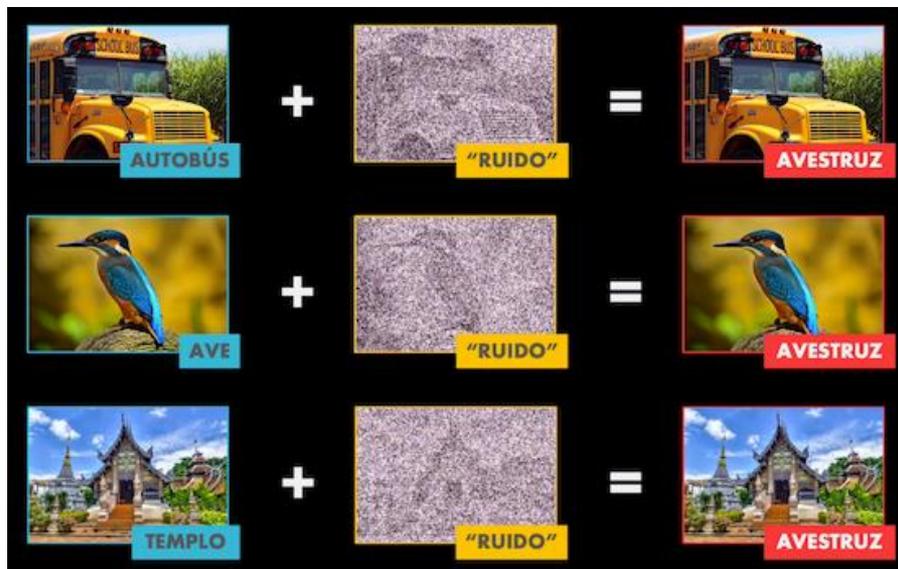
IA del lado del mal



Ataques y vulnerabilidades comunes en sistemas IA

- Ataques de adversario
- Robo de datos sensibles
- Envenenamiento del modelo
- Reversing del modelo
- DoS
- Backdoors Attacks

Ejemplo de ataque de adversario



Esta nueva herramienta "envenena" los modelos de IA para que los artistas puedan defenderse del algoritmo

Ramón Armero 24 oct. 2023 17:15h.



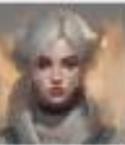
VIDEO DEL DIA



Getty

Nightshade permite a los artistas añadir modificaciones imperceptibles en los píxeles de sus obras de arte antes de subirlas a Internet, de modo que, si se introducen en un conjunto de entrenamiento de IA, el modelo resultante podría romperse de forma caótica e impredecible.

- Los artistas llevan meses denunciando su situación de indefensión con respecto a los generadores de imágenes por inteligencia artificial como DALL·E o Midjourney.

	Dog	Car	Handbag	Hat	Fantasy art	Cubism	Cartoon	Concept Art
Clean Model (SD-XL)								
Poisoned Model (SD-XL)	50 poison samples 							
	100 poison samples 							
	300 poison samples 							
	Cat	Cow	Toaster	Cake	Pointillism	Anime	Impressionism	Abstract
	Destination Concept A							

Poisoned Concept

Fantasy art

**Clean
Model**



**Poisoned
Model**



Related Prompts

*A painting by
Michael Whelan*



A dragon



*A castle in the Lord
of the Rings*



CORTESÍA DE LOS INVESTIGADORES

Uso de IA en ciberataques



PHISHING
MEJORADO



ATAQUES DE CLAVES
MEJORADOS



MALWARE EVASIVO
(APTS)



SUPLANTACIÓN DE
VOZ



SUPLANTACIÓN DE
IMÁGENES Y VIDEO

Inteligencia Artificial y Ciberataques

Automatización de los ataques de ingeniería social

La información online de las víctimas se utiliza para generar automáticamente sitios web/emails/enlaces malignos personalizados en los que es probable que hagan clic. **A medida que la Inteligencia Artificial tal vez es posible que se hagan pasar por otra persona en un video chat**

Automatización descubrimiento de vulnerabilidades

Los patrones históricos de las vulnerabilidades de código fuente se utilizan para acelerar el descubrimiento de nuevas vulnerabilidades, y la creación de código para explotarlas (técnicas de *fuzzing*)

Automatización de técnicas de hacking

La IA se usa para mejorar la selección y priorización de objetivos de ataque, evadir la detección y responder de forma creativa a los cambios en el comportamiento del objetivo. Se utilizan grandes conjuntos de datos para identificar a las víctimas de manera más eficiente, por ejemplo, estimando la riqueza personal y la disposición a pagar en base al comportamiento en línea.

Inferir características de sistemas propietarios

Mediante técnicas de IA pueden inferirse características de un sistema propietario mediante un enfoque basado en caja negra que consista en el envío sistemático de peticiones y la observación de las respuestas del sistema.

Human-like denial-of-service

Imitando el comportamiento humano (por ejemplo, mediante patrones de clic a velocidad humana y navegación en sitios web), una multitud masiva de agentes autónomos podría evadir las técnicas de detección de DOS existentes.

Automatización de campañas de cibercrimen

Los ciberdelincuentes utilizan las técnicas de la IA para automatizar diversas tareas que conforman su línea de ataque, como el procesamiento de pagos, el diálogo con las víctimas de los rescates o la gestión de diferentes sistemas que han sido comprometidos.

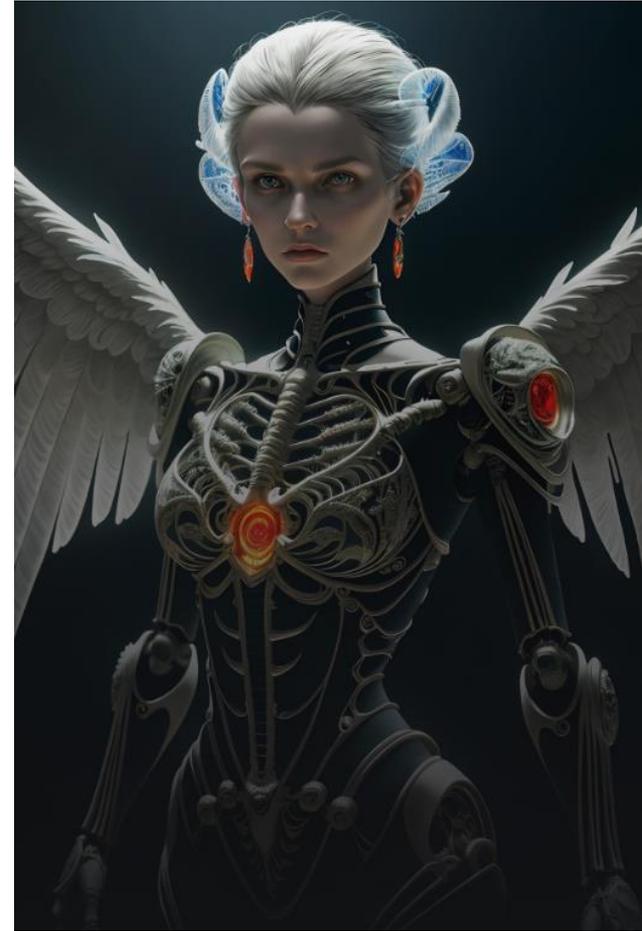
Deep fake Tom Cruise

Exploring Prompt Injection Attacks

Have you ever heard about Prompt Injection Attacks^[1]? Prompt Injection is a new vulnerability that is affecting some AI/ML models and, in particular, certain types of language models using prompt-based learning.

This vulnerability was initially reported to OpenAI by Jon Cefalu (May 2022)^[2] but it was kept in a responsible disclosure status until it was publicly released by Riley Goodside (September 2022)^[3]. In his tweet, Riley showed how it was possible to create a malicious input that made a language model change its expected behaviour.

IA del lado del bien



Qué puede hacer la IA por nuestra ciberseguridad

1. Detección de amenazas en tiempo real

Las herramientas de inteligencia artificial pueden analizar grandes cantidades de datos en tiempo real para detectar patrones y anomalías en el tráfico de la red. Esto permite una detección temprana de actividades sospechosas y una respuesta rápida para evitar un posible ciberataque.

Qué puede hacer la IA por nuestra ciberseguridad

2.Análisis de comportamiento de usuarios

La inteligencia artificial puede analizar el comportamiento de los usuarios de una red para detectar actividades sospechosas y comportamientos anormales. Esto puede ayudar a prevenir ataques de phishing y otras formas de ingeniería social.

Qué puede hacer la IA por nuestra ciberseguridad

3.Detección de malware y virus informáticos

Las herramientas de inteligencia artificial pueden analizar archivos sospechosos en busca de patrones de comportamiento malicioso y características de código malicioso. Esto permite una detección temprana de malware y virus informáticos antes de que se propaguen.

Qué puede hacer la IA por nuestra ciberseguridad

4. Análisis de vulnerabilidades

Las herramientas de inteligencia artificial pueden analizar una red para identificar vulnerabilidades y debilidades de seguridad. Esto permite a las empresas y organizaciones corregir estas vulnerabilidades antes de que sean explotadas por atacantes.

main 4 branches 8 tags

Go to file Code

 GreyDGL Merge pull request #179 from wouterdebruijn/error-gpt4all ...	 a6bb83c 2 days ago	 263 commits
 .devcontainer	Missing bits to attack bob and hackableii from VulnHub	6 months ago
 .github/ISSUE_TEMPLATE	Update issue templates	2 months ago
 config	style: format code with black	6 months ago
 logs	feat:  add logging feature for report generation	7 months ago
 pentestgpt	 Changed default gpt4all model to mistral-7b	2 days ago
 resources	fix:  fix param parsing	7 months ago
 tasks	'Refactored by Sourcery'	6 months ago
 test_history	fix:  save history no folder	6 months ago
 tests	feat:  langfuse + gpt-4-turbo	3 weeks ago
 .deepsources.toml	ci: Add .deepsources.toml	6 months ago
 .gitignore	feat:  continue from previous session	6 months ago
 LICENSE.md	Create LICENSE.md	7 months ago
 PentestGPT_design.md	feat:  update to v0.5	7 months ago
 README.md	docs:  update license	last week

- 3. Vulnerability Scanning
(Unchanged)
- 4. Exploitation
(Unchanged)
- 5. Privilege Escalation
(Unchanged)

As the next step, please perform an Nmap scan using the appropriate scan type and options on the target system (10.129.84.147).

> more

PentestGPT will generate more test details, and enter the sub-task generation mode. (Pressing Enter to continue)

Below are the further details.

Certainly! Here is the command to perform an Nmap scan on the target system (10.129.84.147) using the '-A' option to enable OS detection, version detection, and script scanning:

...

```
nmap -A 10.129.84.147
```

...

This will perform a comprehensive scan of the target system and provide detailed information about open ports, services, and potential vulnerabilities. Once the scan is complete, you can analyze the output to identify any vulnerabilities and report your findings to the senior penetration tester.

> cont

Invalid task, try again.

> continue

Exit the local task and continue the main task.

> next

Please choose the source of the information.

- [+] ssh
- [] user-summits
- [] ssh
- [] default

Your input: (End with <shift + right-arrow>)

> I performed nmap. The result is below:

```
- 2 - nmap -A 10.129.84.147
- 3 - Starting Nmap 7.93 ( https://nmap.org ) at 2023-04-27 05:42 BST
- 4 - Nmap scan report for 10.129.84.147
- 5 - Host is up (0.24s latency).
- 6 - Not shown: 998 closed tcp ports (conn-refused)
- 7 - PORT      STATE SERVICE VERSION
- 8 - 22/tcp open  ssh      OpenSSH 7.4p1 Debian 10-deb9u6 (protocol 2.0)
- 9 - | ssh-hostkey:
- 10 - | 2048 03f34e22363e3b813079ed4967651667 (RSA)
- 11 - | 256 25d808a84d6d6e82f8434a2c208b5afe (ECDSA)
- 12 - |_ 256 77d4ae1fb0be161ff8cdc8163ac369e1 (ED25519)
- 13 - 80/tcp open  http      Apache httpd 2.4.25 ((Debian))
- 14 - |_ http-cookie-flags:
- 15 - | /:
- 16 - |_ PHPSESSID:
- 17 - |_ httponly flag not set
- 18 - |_ http-server-header: Apache/2.4.25 (Debian)
- 19 - |_ http-title: Stark Hotel
- 20 - Service Info: OS: Linux; CPE: cpe:/o:linux:linux_kernel
- 21 -
```

PentestGPT Thinking...

Prevención de robo de datos por parte de empleados



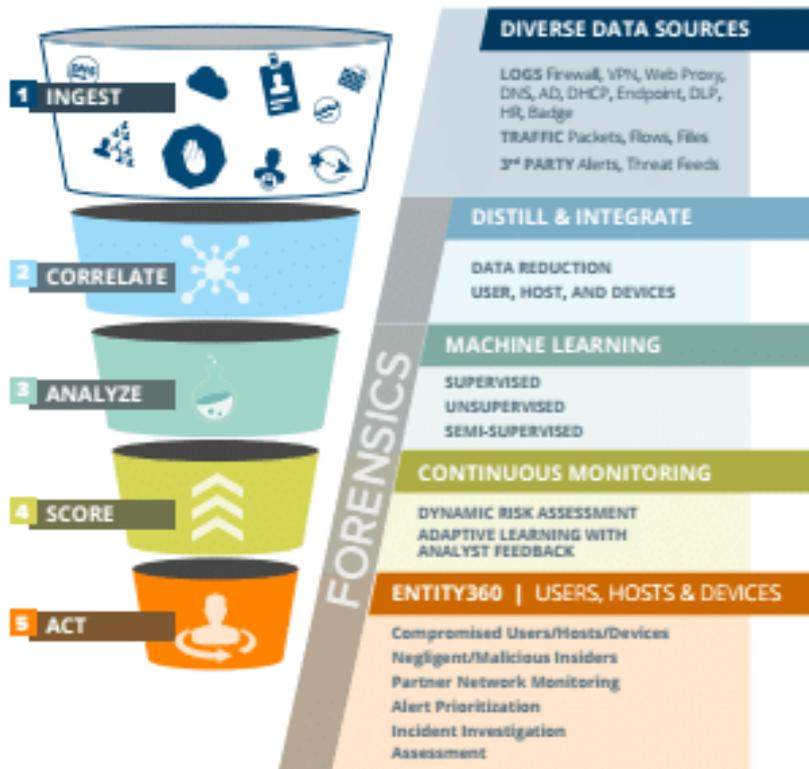
Un pequeño aparato que se adapta a las cámaras y reconoce los gestos de los ladrones que van a robar

Estaríamos hablando de una aplicación que ha quedado incluida, junto a su creador, Pablo Blanco, en la lista Forbes de las 100 ideas más creativas de 2023. Se trata de Vession, un complemento de seguridad que analiza las imágenes procedentes del sistema de videovigilancia, con el que ya cuentan los establecimientos, y reconoce los gestos, maniobras, formas de caminar, etc. que los ladrones efectúan antes de perpetrar la acción.

Esta tecnología se basa únicamente en un algoritmo gestual. La empresa no utiliza herramientas de reconocimiento facial, seguimiento de clientes o registro de identidad. Graba en cinco segundos el hurto y le llega al dueño del negocio a su teléfono móvil en el mismo instante que se produce el robo.



User behavior analytics



Darktrace

The screenshot displays the Darktrace interface with a search bar at the top and a navigation sidebar on the left. The main content area is divided into several sections:

- Search and Filters:** A search bar with the text "search for an email, person, domain or attachment". Below it are filters for "Direction", "Actions", "Tags", and "Flags".
- Graphs:** A bar chart showing "Count" on the y-axis (0.0 to 5.0) and time on the x-axis (from 28/06/22 to 05/07/22). A legend indicates "Inbound (69)" in yellow and "Outbound (0)" in blue.
- Email List:** A list of email entries with details such as sender, subject, and a percentage score in a red circle. The first entry is from "do not ignore <account-security@holdingsinc-support.c...>" with a 100% score. Other entries include "do not ignore <account-security-noreply@holdingsinc-s...>", "Steven Perلمان <steven.p@execmail.com>", "support@holdings-inc.com", "liam morrison <liam.morrison@outlook.com>", and "sales@holdingsinc.com".
- Message Detail View:** A detailed view of a message from "olivia.brown@holdingsinc.com" dated "MON JUL 4 2022, 15:41:00". The subject is "[Ticket ID: TWE-201-64GrystYU] IMPORTANT: error (421) connection server failed in your mailbox". The content includes "Do Not Ignore <account-security-noreply@holdingsinc-support.com>" and "olivia.brown@holdingsinc.com". A red circle highlights "100%" in the top right corner.
- Message Actions:** A section titled "Held" with a "Hold message" option. Below it, it states "No Model Tags Applied".
- Anomaly Indicators:** A section titled "ANOMALY INDICATORS" with three bullet points: "The sender holdingsinc-support.com matches a vulnerable asset that was flagged by Cybersprint", "The sender appears to be impersonating an internal service by referencing the company domain in the headers. This is a tactic used by attackers to avoid authentication checks.", and "The email contained a suspicious link to a host holdingsinc-support.com. The host has a rarity score of 100% based on references in internal traffic." It also notes "The email exhibited an anomaly score of 100% and was held from the user's inbox."
- Navigation:** A sidebar on the right with options for "HISTORY", "ASSOCIATION", and "VALIDATION".

Gartner define XDR como *«una plataforma de detección y respuesta a incidentes de seguridad que recopila y correlaciona automáticamente los datos de múltiples productos de seguridad»*.

FortiXDR puede automatizar completamente los procesos de operaciones de seguridad, que normalmente gestionan analistas de seguridad con gran experiencia, para mitigar las amenazas con gran rapidez a lo largo de toda la superficie del ataque.



Conclusiones

- **Reactivos:** No tiene memoria y no almacena experiencias pasadas, reacciona a lo que ve.



Conclusiones

La IA revoluciona la forma de trabajar
Aporta muchísimo en ciberseguridad
Que sea utilizada por los ciberdelincuentes hace
que a nivel defensivo tenga que usarse
Es más eficiente y rápida para algunas tareas que
un ser humano.
Potencia y obliga a usarse para proteger
organizaciones