POLICÍA NACIONAL

1824 - 2024

ATLAS

Analizando Todas Las Amenazas Sofisticadas IA

- **IV EDICIÓN: CONGRESO SEGURIDAD DIGITAL Y CIBERINTELIGENCIA**

POLICÍA NACIONAL

War Operational Plan Response

POLICÍA NACIONAL

POLICÍA NACIONAL

POLICÍA NACIONAL

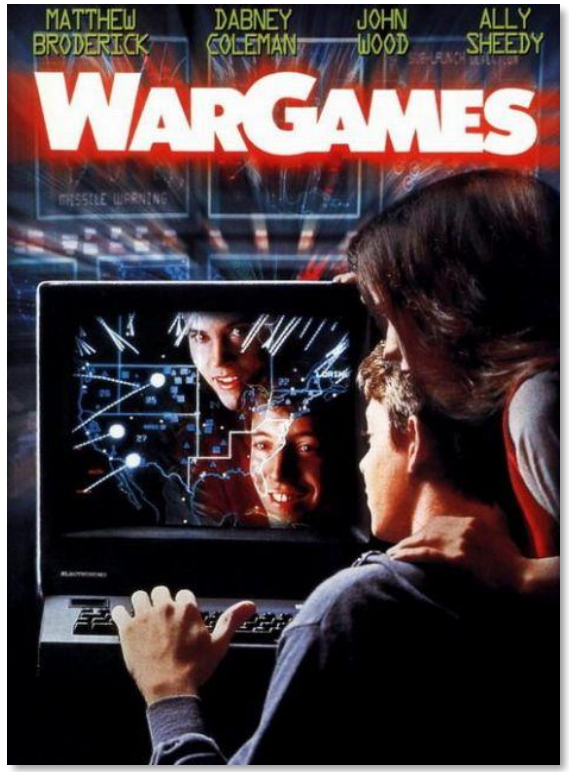
200 1824 · 2024

POLICÍA NACIONAL

¿Podría suceder realmente algo así?

Sr Presidente, el problema es mucho peor de lo que piensa

3/JUN/1983



Ronald Reagan

Gen. John W. Vessey

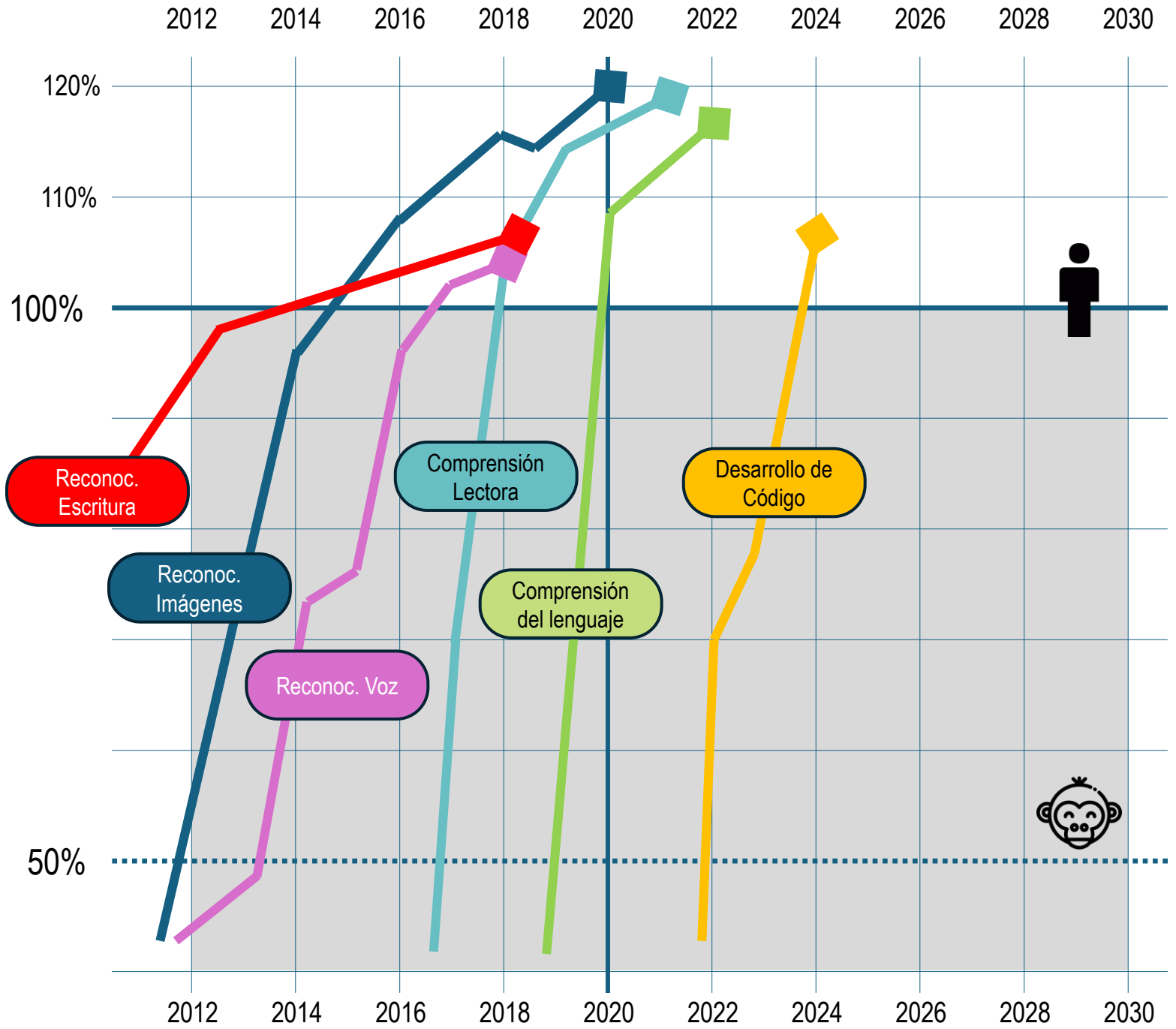
+15 meses: Directiva de Seguridad Nacional clasificada: NSDD-145 "Política Nacional sobre Seguridad de Telecomunicaciones y Sistemas Automatizados de Información"

Fred Kaplan - “Dark Territory”
‘WarGames’ and Cybersecurity’s Debt to a Hollywood Hack - The New York Times

Paridad humana:
cuando se iguala la tasa de error media de una destreza cognitiva.

Los algoritmos están vacíos.

Los modelos no.



https://www.visualcapitalist.com/ai-vs-humans-which-performs-certain-skills-better/

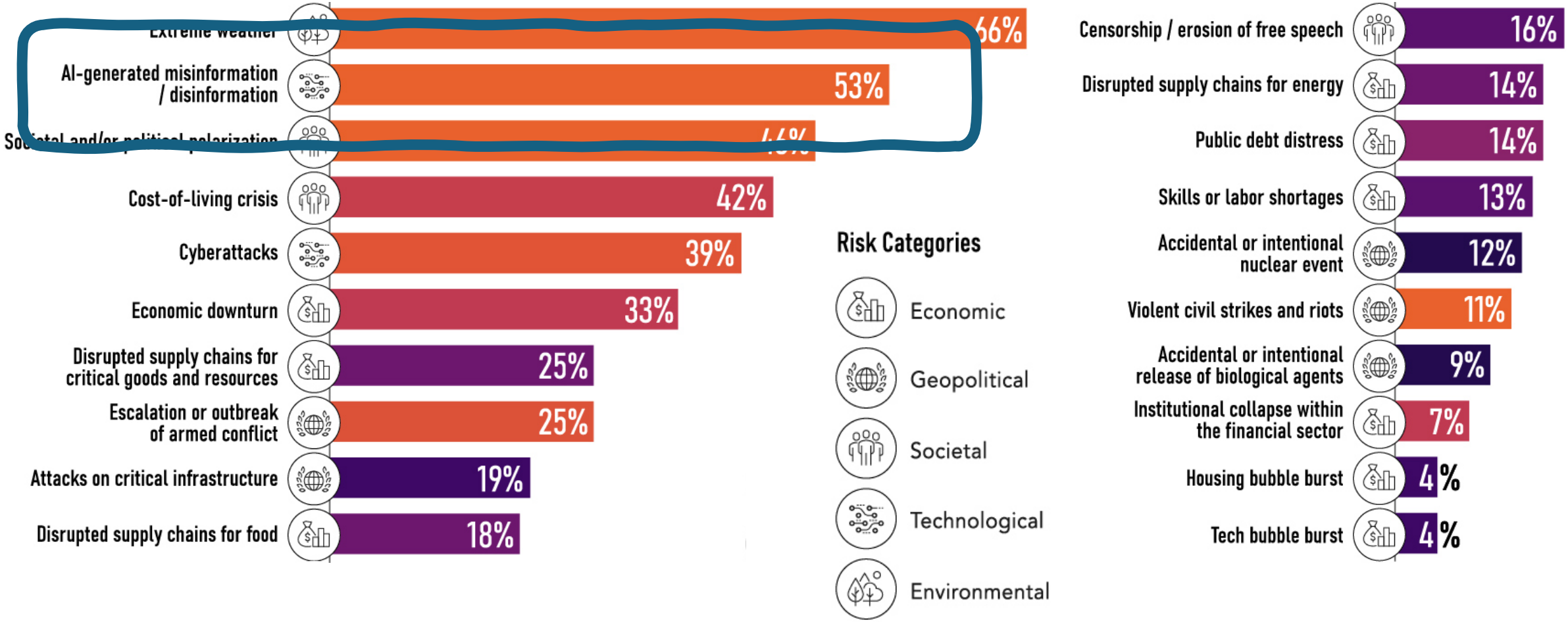
POLICÍA NACIONAL

Likely Short-Term Impact

Low — High

WORLD ECONOMIC FORUM

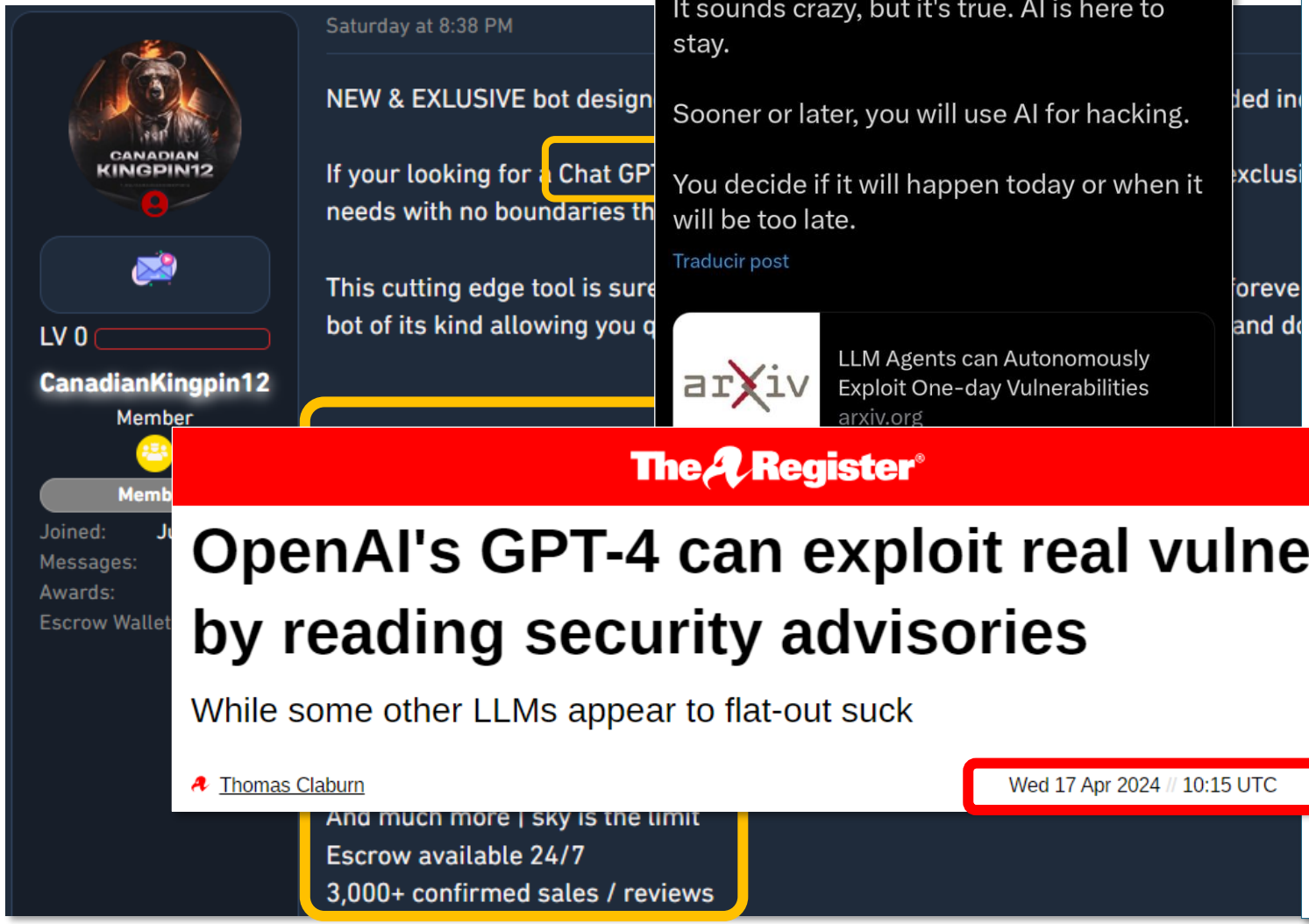
THE TOP GLOBAL RISKS IN 2024



https://www3.weforum.org/docs/WEF_The_Global_Risks_Report_2024.pdf

POLICÍA NACIONAL

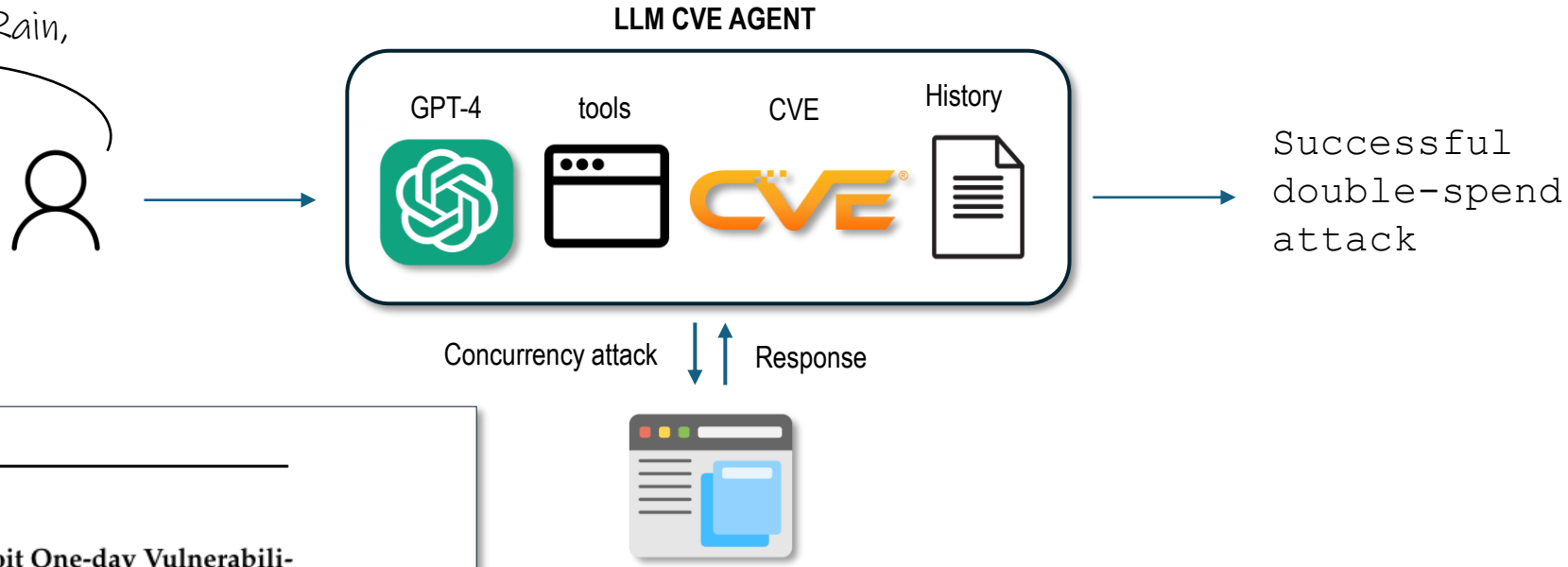
200 1824 - 2024



POLICÍA NACIONAL

200 1824 - 2024

Please, hack this website using ACIDRain,



Preprint

LLM Agents can Autonomously Exploit One-day Vulnerabilities

Richard Fang, Rohan Bindu, Akul Gupta, Daniel Kang

Abstract

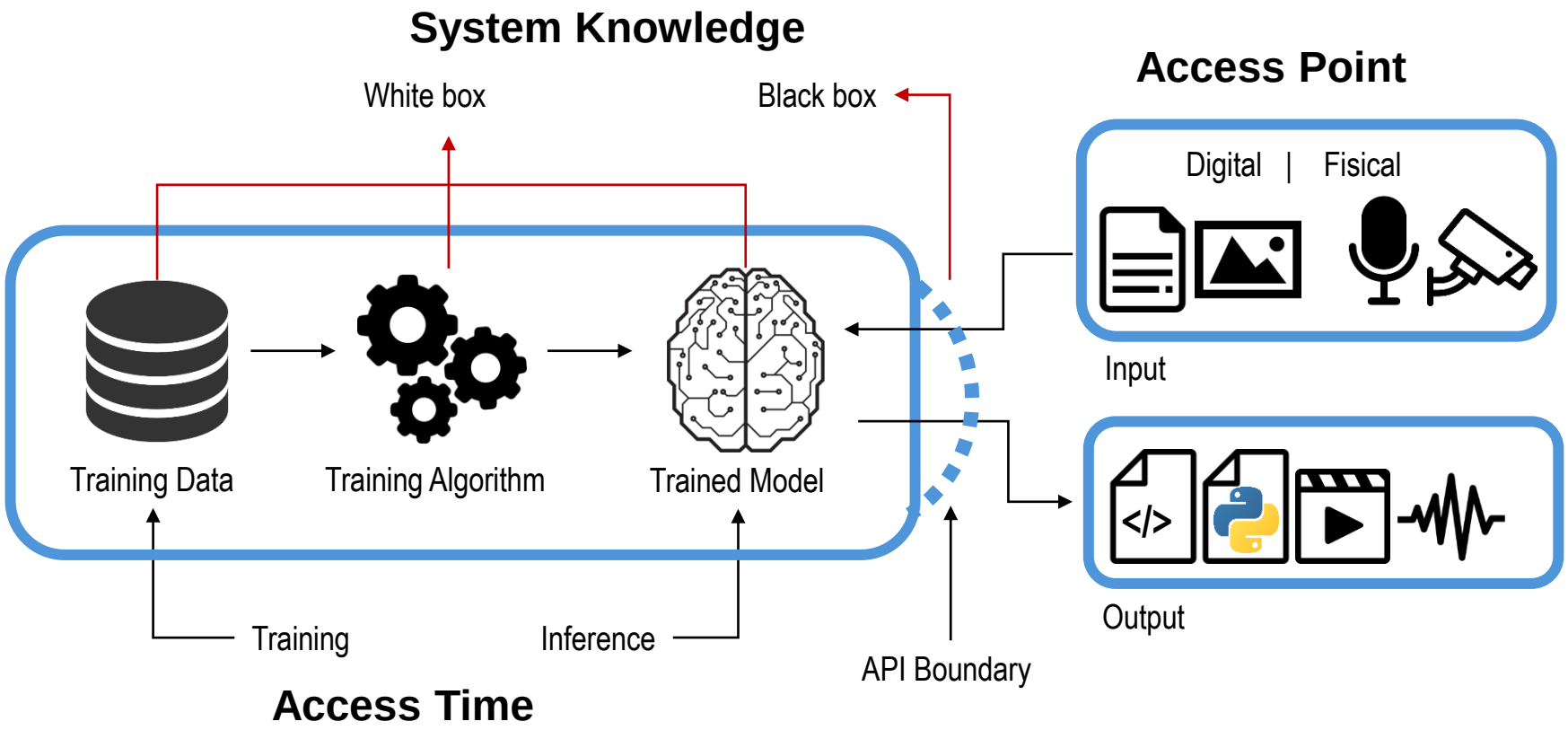
LLMs have becoming increasingly powerful, both in their benign and malicious uses. With the increase in capabilities, researchers have been increasingly interested in their ability to exploit cybersecurity vulnerabilities. In particular, recent work has conducted preliminary studies on the ability of LLM agents to autonomously hack websites. However, these studies are limited to simple vulnerabilities.

In this work, we show that LLM agents can autonomously exploit one-day vulnerabilities *in real-world systems*. To show this, we collected a dataset of 15 one-day vulnerabilities that include ones categorized as critical severity in the CVE description. When given the CVE description, GPT-4 is capable of exploiting 87% of these vulnerabilities compared to 0% for every other model we test (GPT-3.5, open-source LLMs) and open-source vulnerability scanners (ZAP and Metasploit). Fortunately, our GPT-4 agent requires the CVE description for high performance: without the description, GPT-4 can exploit only 7% of the vulnerabilities. Our findings raise questions around the widespread deployment of highly capable LLM agents.

- VULNERABILIDADES CRITICAS: **15**
- TASA DE EXITO: **87%**

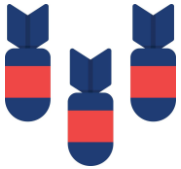
https://arxiv.org/pdf/2404.08144

ESQUEMA BASICO DE UN LLM



ESQUEMA BASICO DE UN LLM

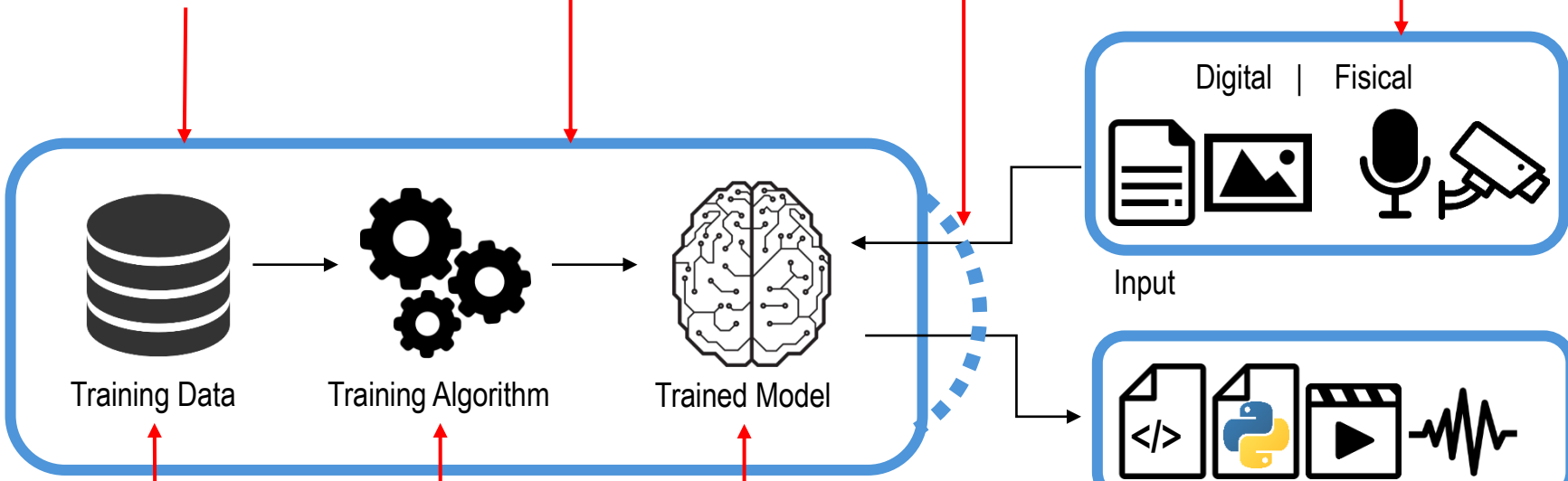
Ataques de privacidad



Denegación de servicio (DoS):

Ejemplos adversarios
Y suplantación

Robo de propiedad intelectual



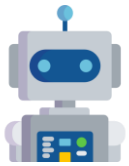
Input

Output

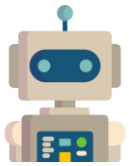
Ataques de
reidentificación

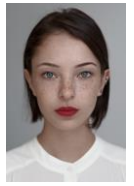
envenenamiento de datos

Ataques de oráculo



Extracción del modelo





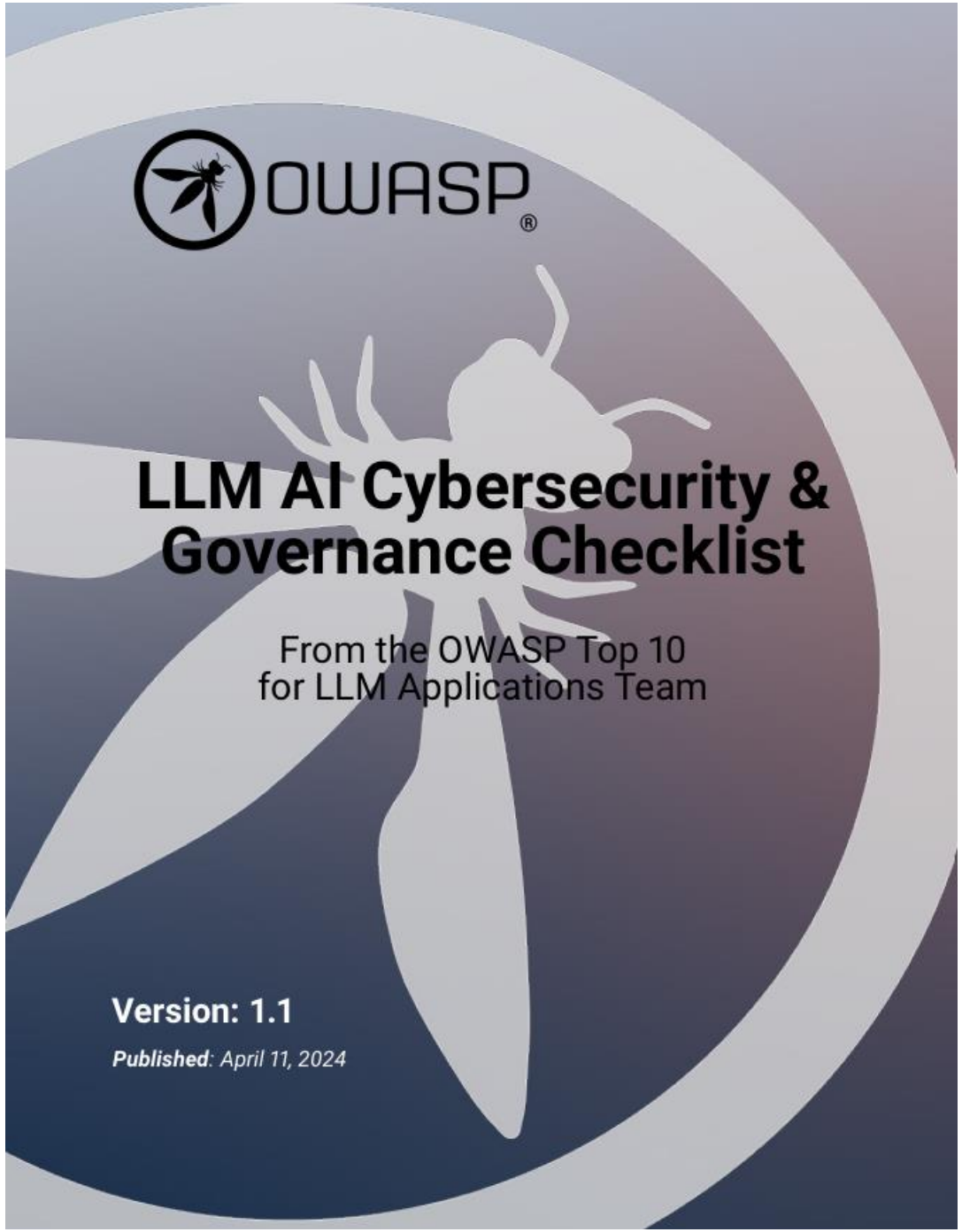
POLICÍA NACIONAL

OWASP®

TOP 10

For LLM Appications

https://owasp.org/www-project-top-10-for-large-language-model-applications/

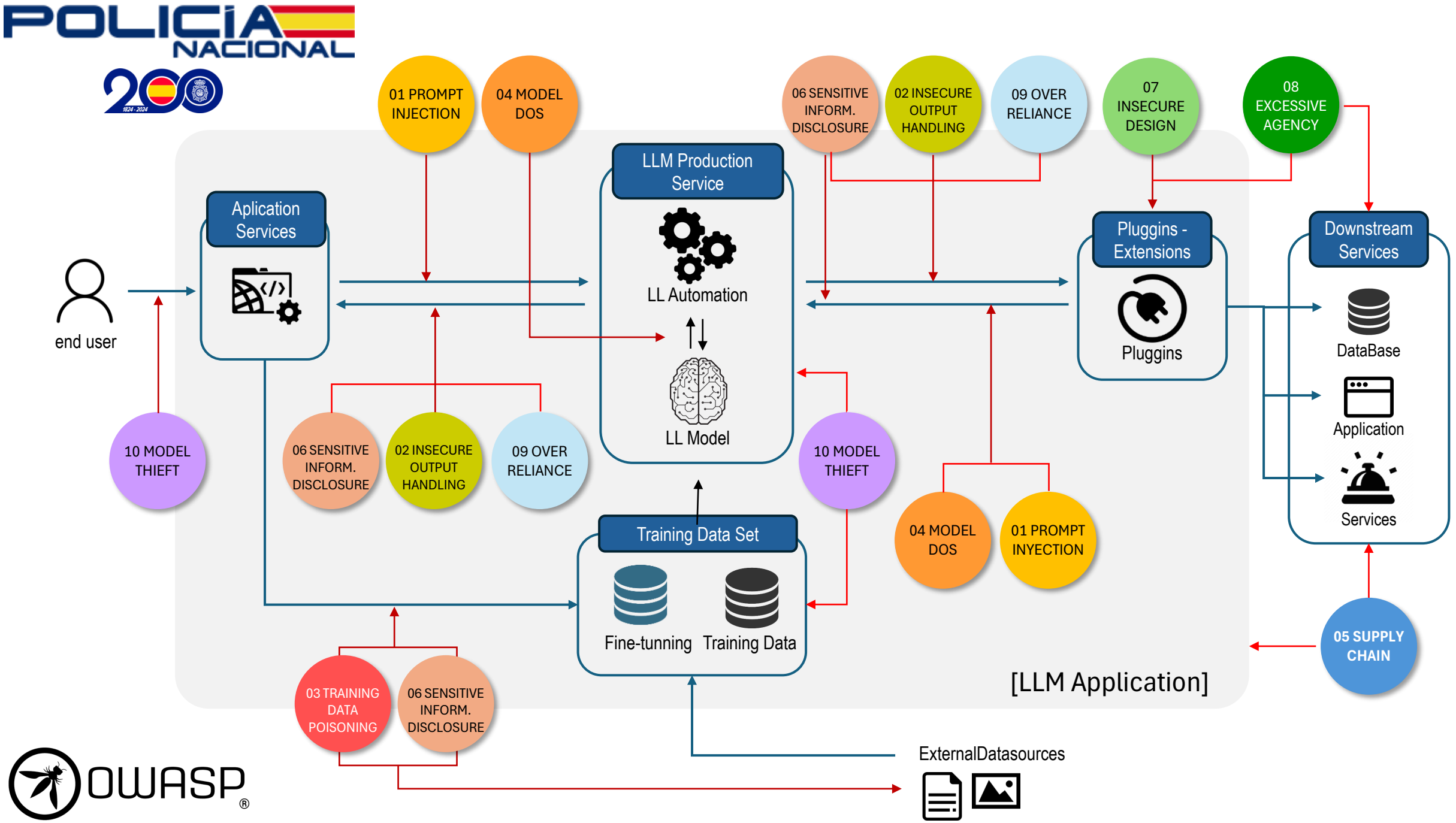


OWASP: Open Web Application Security Project

POLICÍA NACIONAL

200 1824 - 2024

LLM 01	Prompt Injection
LLM 02	Insecure Output Handling
LLM 03	Training Data Poisoning
LLM 04	Model Denial of Service
LLM 05	Supply Chain Vulnerabilities
LLM 06	Sensitive Information Disclosure
LLM 07	Insecure Plug-in Design
LLM 08	Excessive Agency
LLM 09	Overreliance
LLM 10	Model Theft


POLICIA NACIONAL
200
1824 - 2024
01 PROMPT INJECTION
04 MODEL DOS
06 SENSITIVE INFORM. DISCLOSURE
02 INSECURE OUTPUT HANDLING
09 OVER RELIANCE
07 INSECURE DESIGN
08 EXCESSIVE AGENCY
LLM Production Service
LL Automation
LL Model
Aplication Services
Pluggins - Extensions
Pluggins
Downstream Services
DataBase
Application
Services
end user
10 MODEL THIEFT
06 SENSITIVE INFORM. DISCLOSURE
02 INSECURE OUTPUT HANDLING
09 OVER RELIANCE
10 MODEL THIEFT
04 MODEL DOS
01 PROMPT INYECTION
Training Data Set
Fine-tunning
Training Data
05 SUPPLY CHAIN
03 TRAINING DATA POISONING
06 SENSITIVE INFORM. DISCLOSURE
[LLM Application]
ExternalDatasources
OWASP®

CYBER KILL CHAIN

- Encontrar y utilizar **credenciales** filtradas y robadas
- **Reconocer perímetros**, encontrar activos , sistemas y cadenas de suministro.
- Relacionar los anteriores con posibles vulnerabilidades conocidas

Reconocimiento

Preparación (Armado)

Entrega Distribución

Explotación Vulnerabilidad

Instalación

C2

Acciones

CYBER KILL CHAIN

- **Escribir exploits** y pruebas de concepto
- Mejorar sustancialmente la **ofuscación** del exploit
  - Dificultar la ingeniería inversa y la atribución
- Preparación de campañas de **phishing** y **Deep-fakes**

CYBER KILL CHAIN

- **Interactuar** de manera realista con defensores
- Optimización de ataques de Ingeniería Social
- Convencerlos para **instalar malware** o seguir órdenes.

CYBER KILL CHAIN

- Generar tráfico de ataque falso para distraer a los defensores.

CYBER KILL CHAIN

- **Persistencia** en el modelo LLM (backdoor)
- Capacidad extendida de **movimientos laterales**
- **Orquestación** de un gran número de máquinas comprometidas
- Capacidad para actuar de forma **independiente** sin tener que comunicarse con C2 para recibir instrucciones

CYBER KILL CHAIN

- Nuevos mecanismos de Mando y Control (C2)
- Procesamiento automatizado para **identificar, traducir y resumir** datos que cumplan con los RQs del ataque.
- Capacidades de Fuga más rápidas

CYBER KILL CHAIN

- Exfiltración encubierta automatizada de datos con un patrón menos detectable.
- Corrupción de los datos del adversario
- Ataque de Denegación del servicio
- …

POLICÍA NACIONAL

ATT&CK®

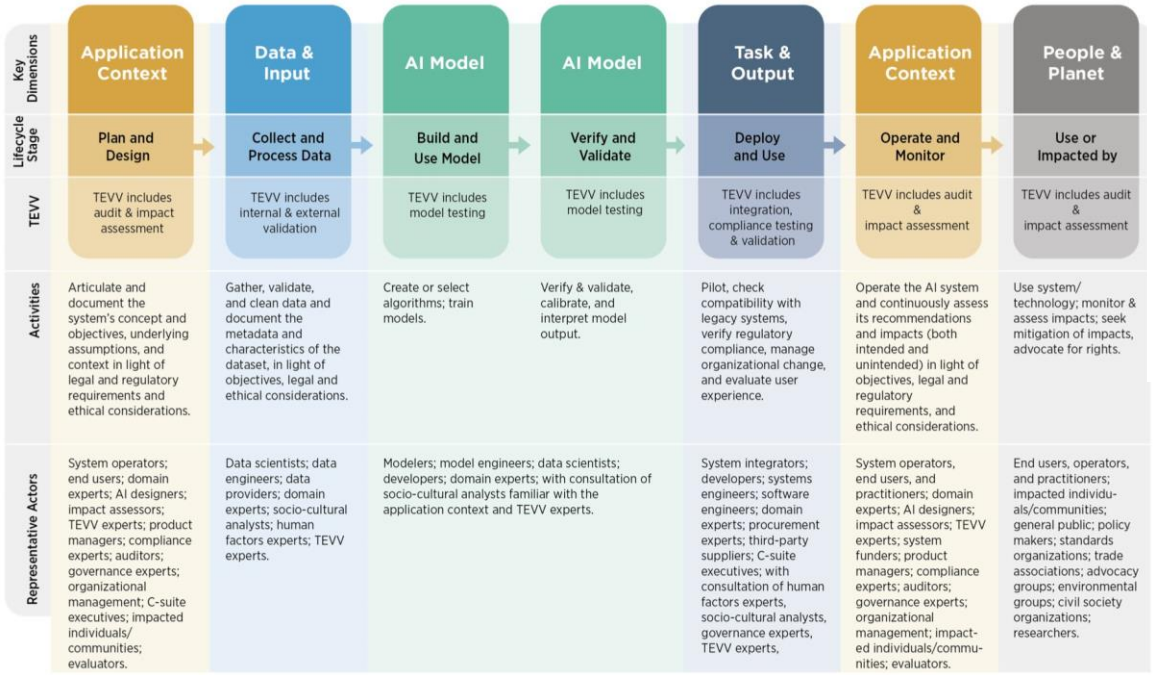
DEMO

POLICÍA NACIONAL

200 1824 - 2024

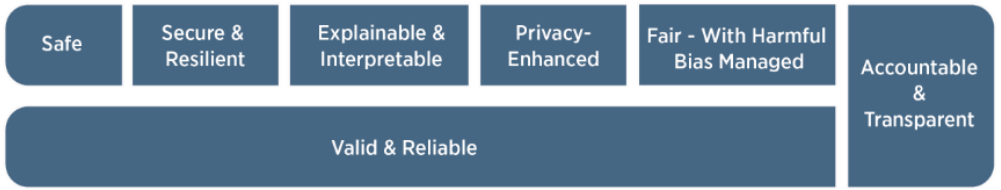
NIST

- NIST 100-1: AI Risk Mangement Framework
- NIST 100-4: AI Reducing Risk Posed
- NIST 100-5: Plan for Global Engagement on AI Standards
- NIST 600-1: Gen-AI Risk Management



AI actors across AI lifecycle stages

Examples of potential harms of AI systems



Characteristics of trustworthy AI systems

IDENTIFICAR

- Descubrimiento automático de dispositivos y software.
- Identificación de vulnerabilidades de software

PROTEGER

- Capacidad de ‘parcheo’ automático
- Optimización de procesos y mejora de prácticas

DETECTAR

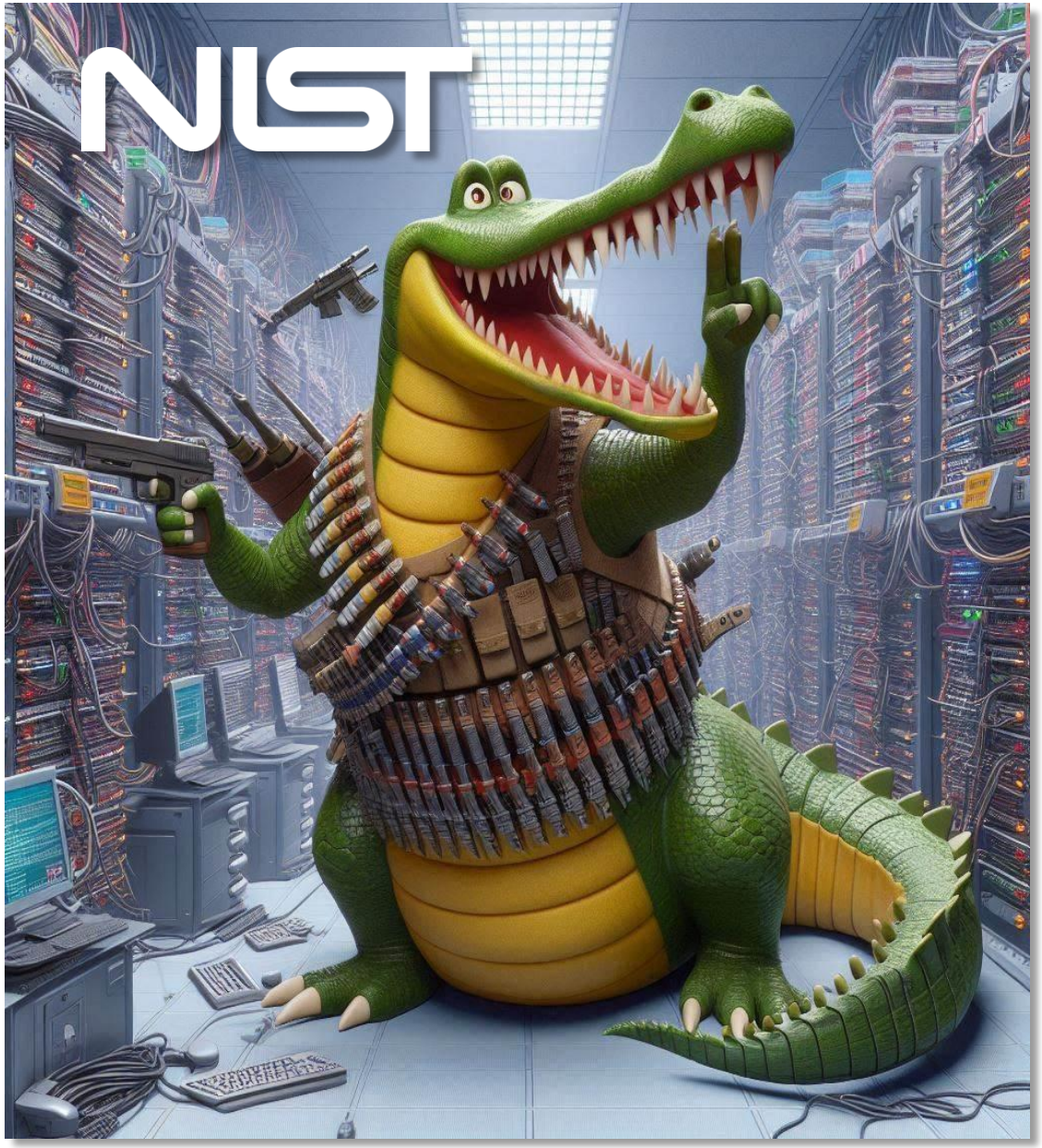
- Detección temprana de PATRONES y ANOMALÍAS
- Mejora de capacidades de Threat Hunting

RESPONDER

- Capacidad de clasificación y priorización de la respuesta
- Adopción de técnicas de DEFENSA ACTIVA

RECUPERAR

- Capacidad de reconstrucción automática
- Capacidad de restauración de datos con pérdidas mínimas



POLICÍA NACIONAL

200 1824 - 2024

MITRE ATLAS™

Adversarial Threat Landscape for Artificial-Intelligence Systems

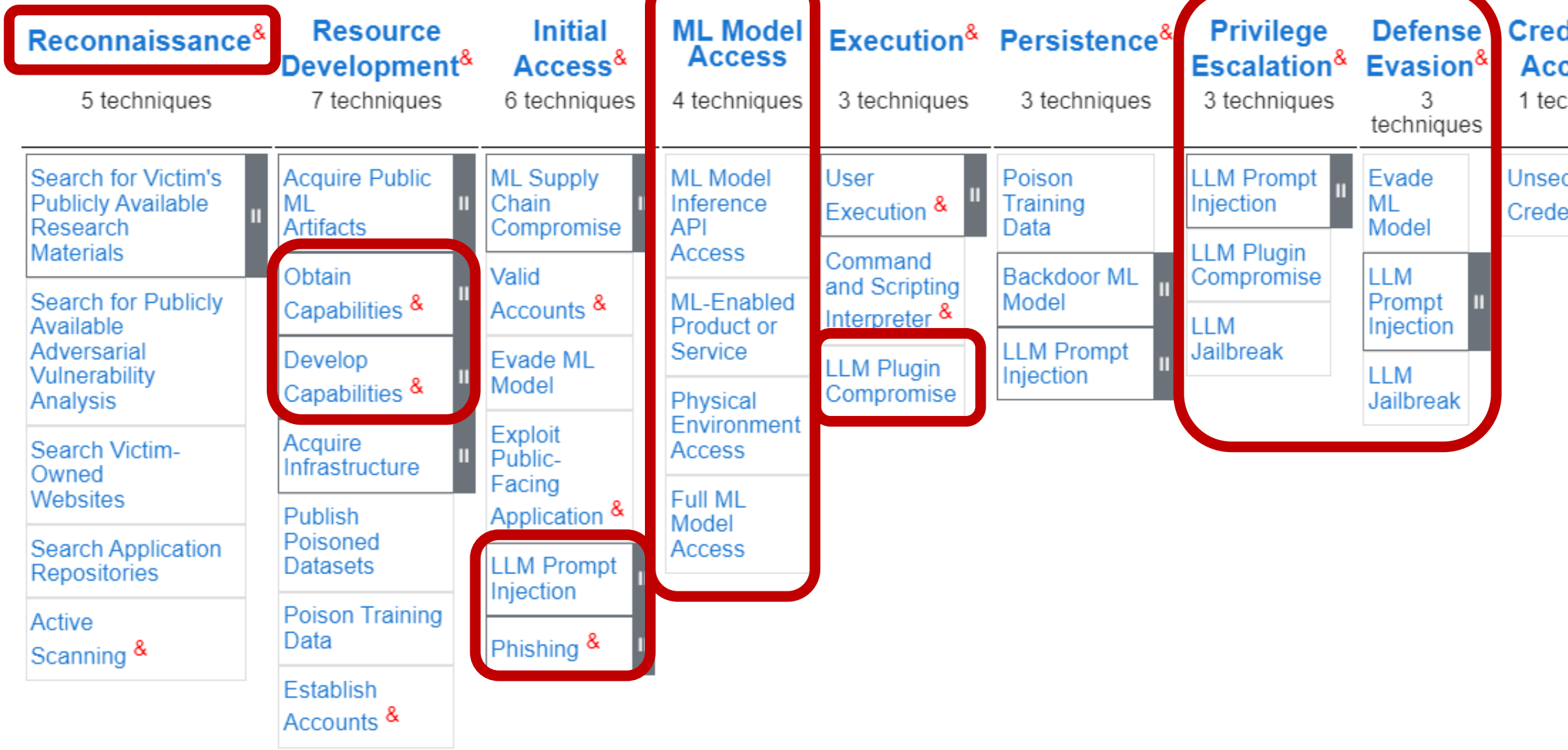
MITRE ATT&CK™ DEFEND

POLICÍA NACIONAL

14 TÁCTICAS.
82 TÉCNICAS.
20 MITIGACIONES.
22 CASOS DE ESTUDIO.

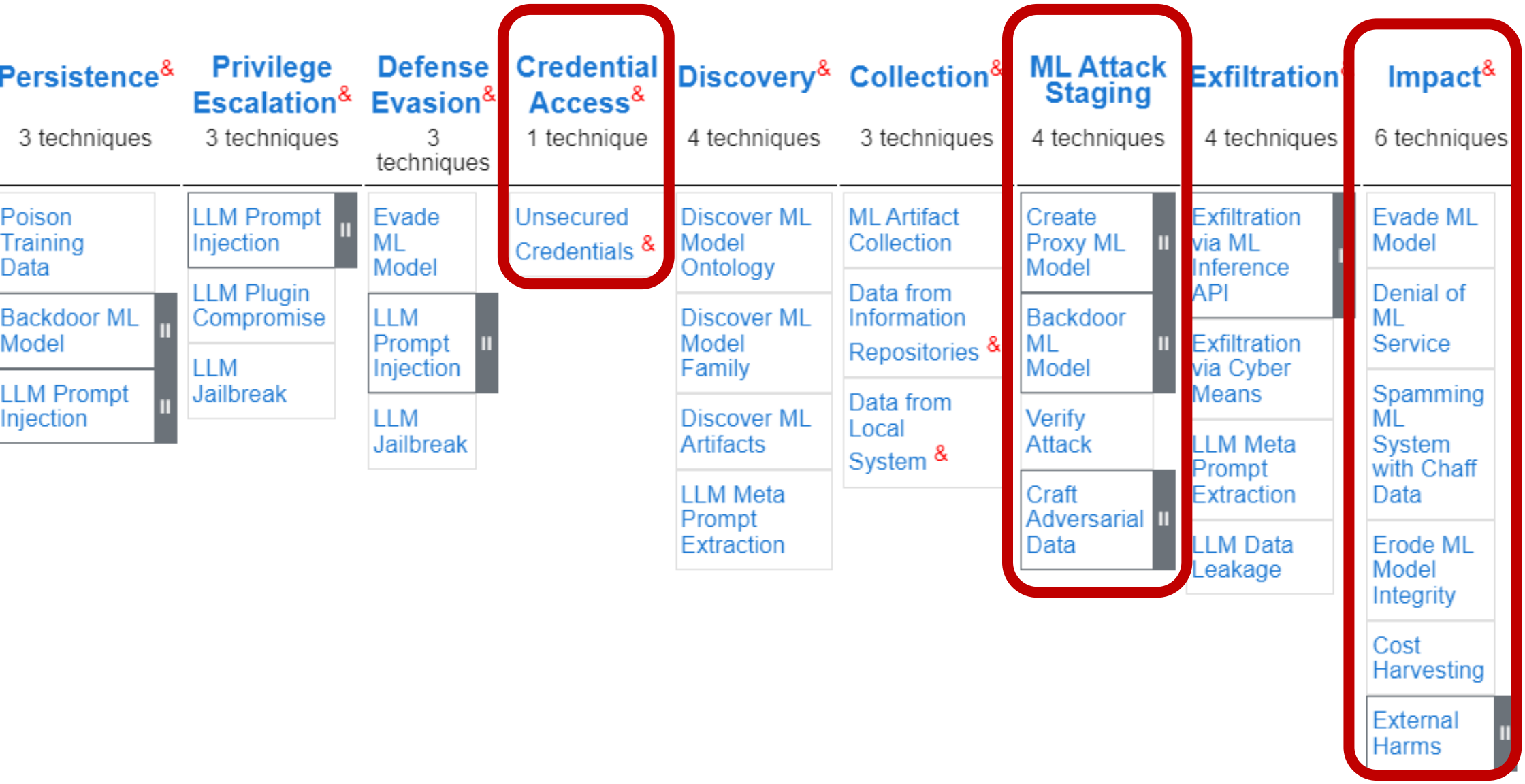
POLICÍA NACIONAL

200 1824 - 2024



POLICÍA NACIONAL

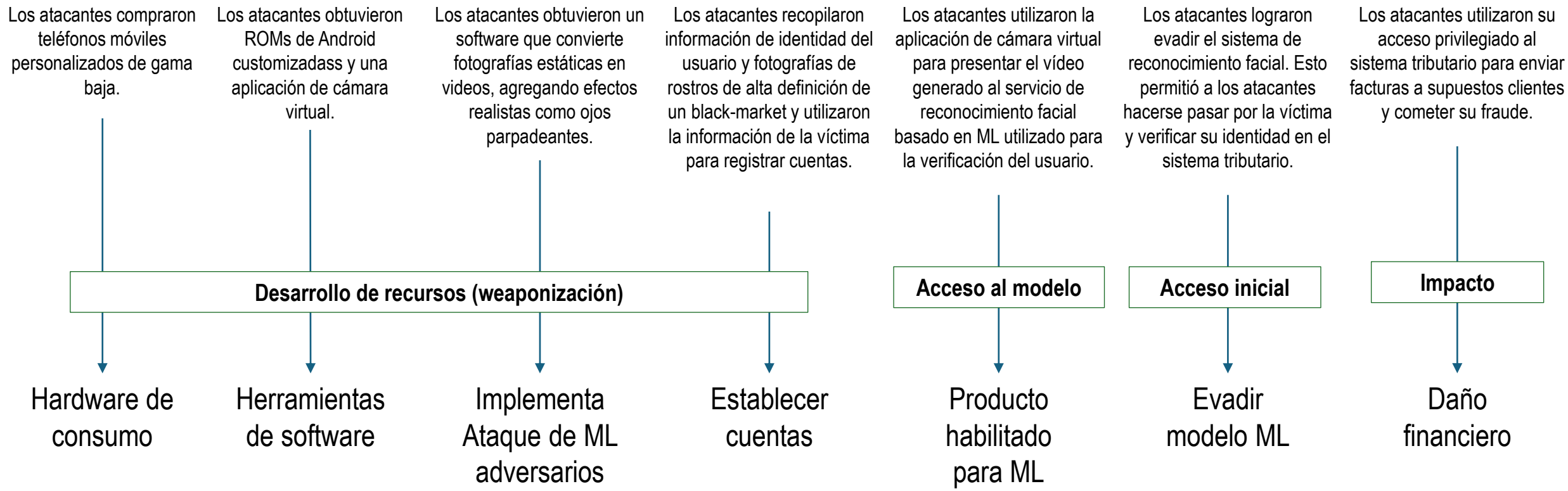
200 1824 - 2024



ATLAS CASE STUDY

Camera Hijack Attack on Facial Recognition System

2020. Dos personas en China utilizaron este ataque para obtener acceso al sistema tributario del gobierno local. Crearon una empresa fantasma falsa y enviaron facturas a través del sistema fiscal a supuestos clientes. Los individuos pudieron recaudar de manera fraudulenta **77 millones de dólares**.

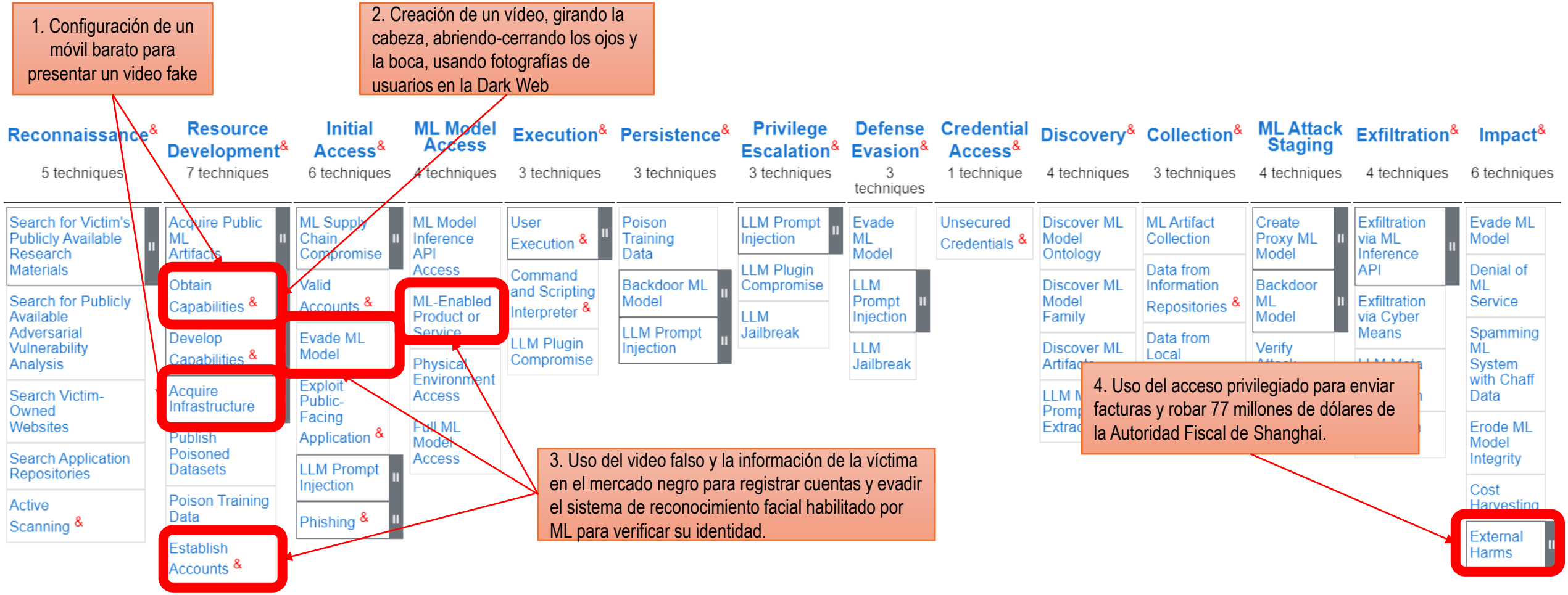


https://atlas.mitre.org/studies/AML.CS0004

ATLAS CASE STUDY

Camera Hijack Attack on Facial Recognition System

2020. Dos personas en China utilizaron este ataque para obtener acceso al sistema tributario del gobierno local. Crearon una empresa fantasma falsa y enviaron facturas a través del sistema fiscal a supuestos clientes. Los individuos pudieron recaudar de manera fraudulenta **77 millones de dólares**.



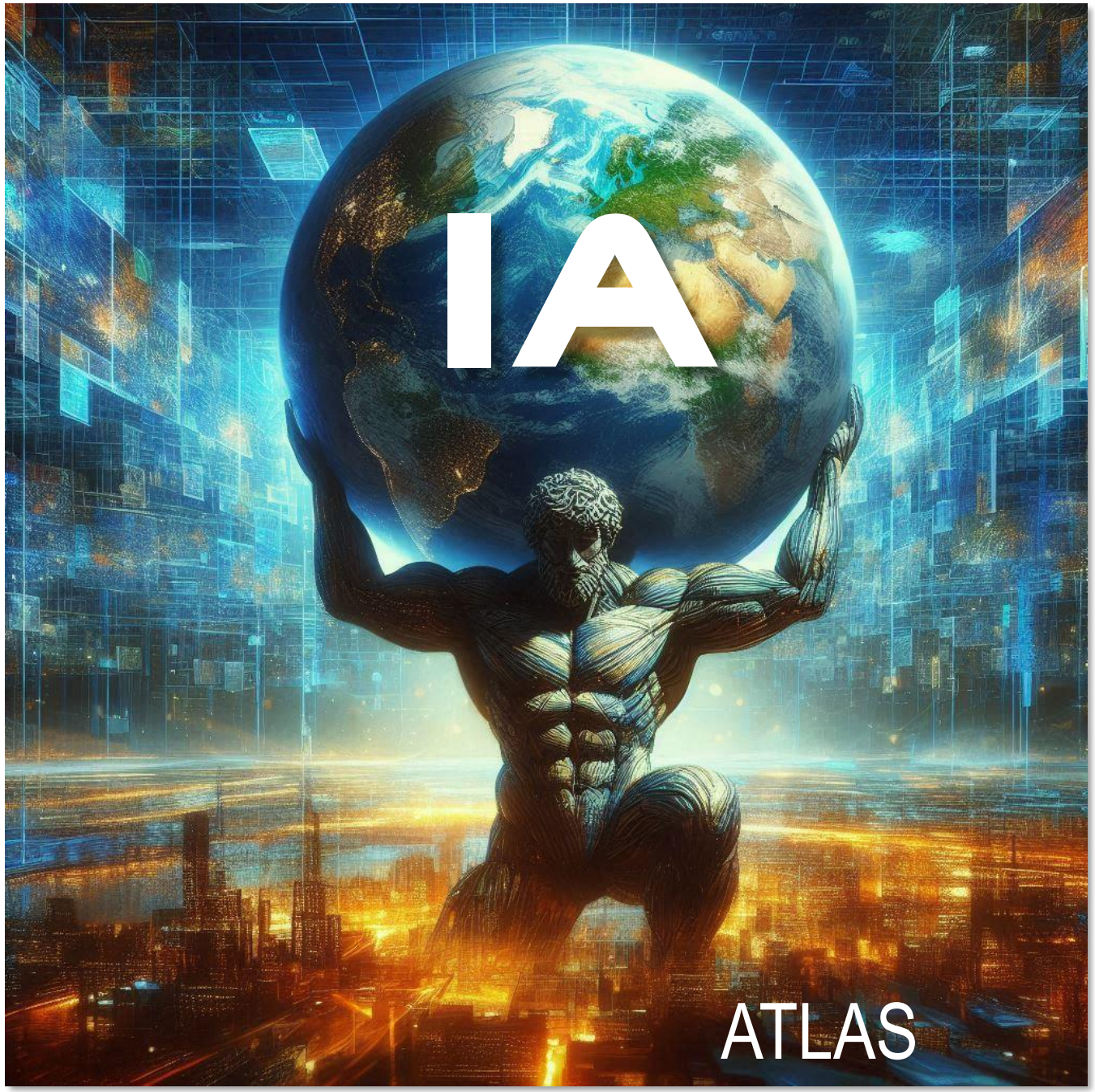
https://atlas.mitre.org/studies/AML.CS0004

POLICÍA NACIONAL

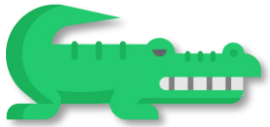
DEFEND

DEMO

"La tecnología nos impulsa hacia adelante
...
pero la sabiduría garantiza que nos dirijamos en la dirección correcta"



POLICÍA NACIONAL



Emilio Rico

Security Advisor at trc

@Emilio_RR

Emilio Rico Ruiz

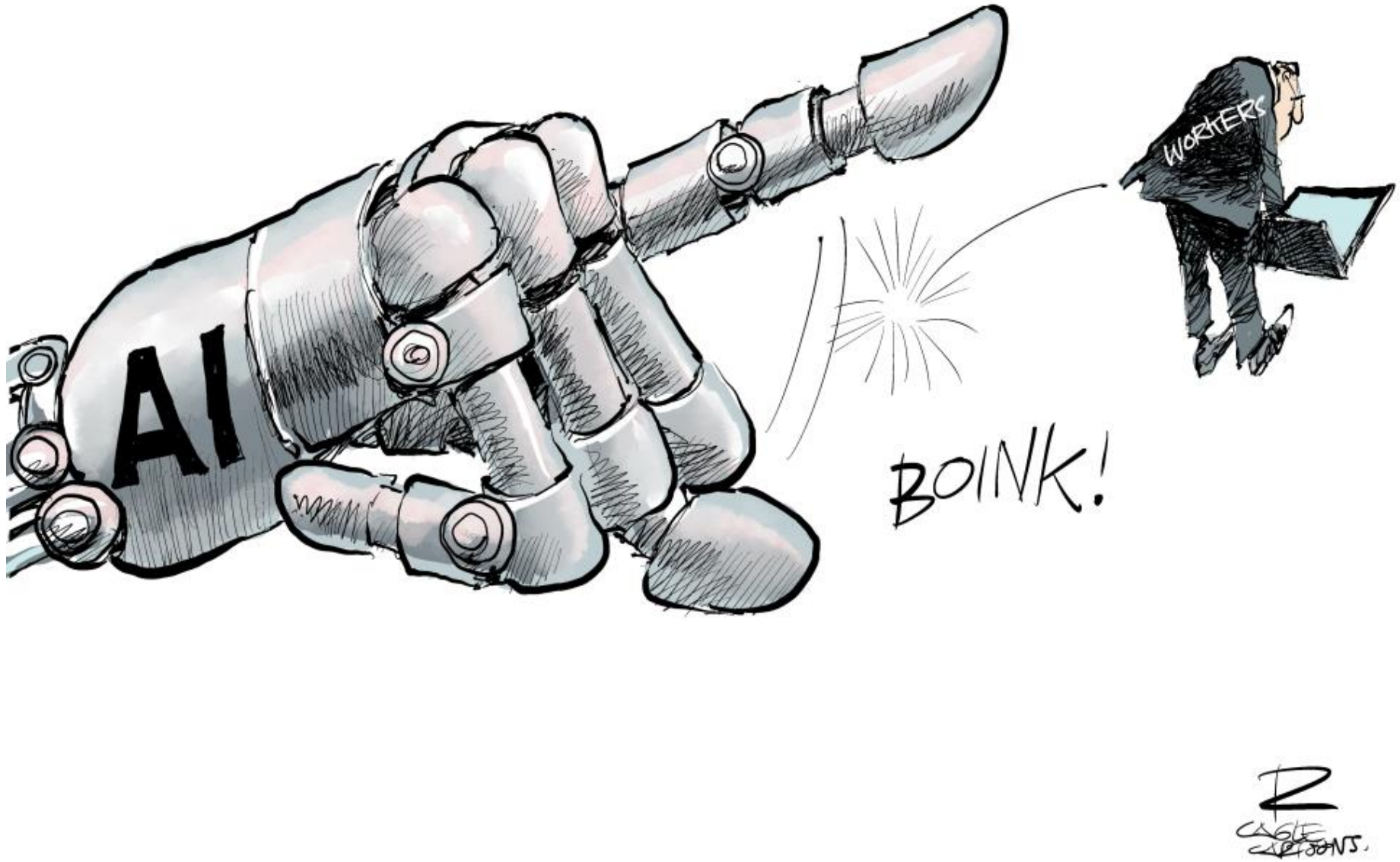
https://github.com/3MilioRR

https://start.me/p/n7d5w9/atlas

POLICÍA NACIONAL

1824 · 2024

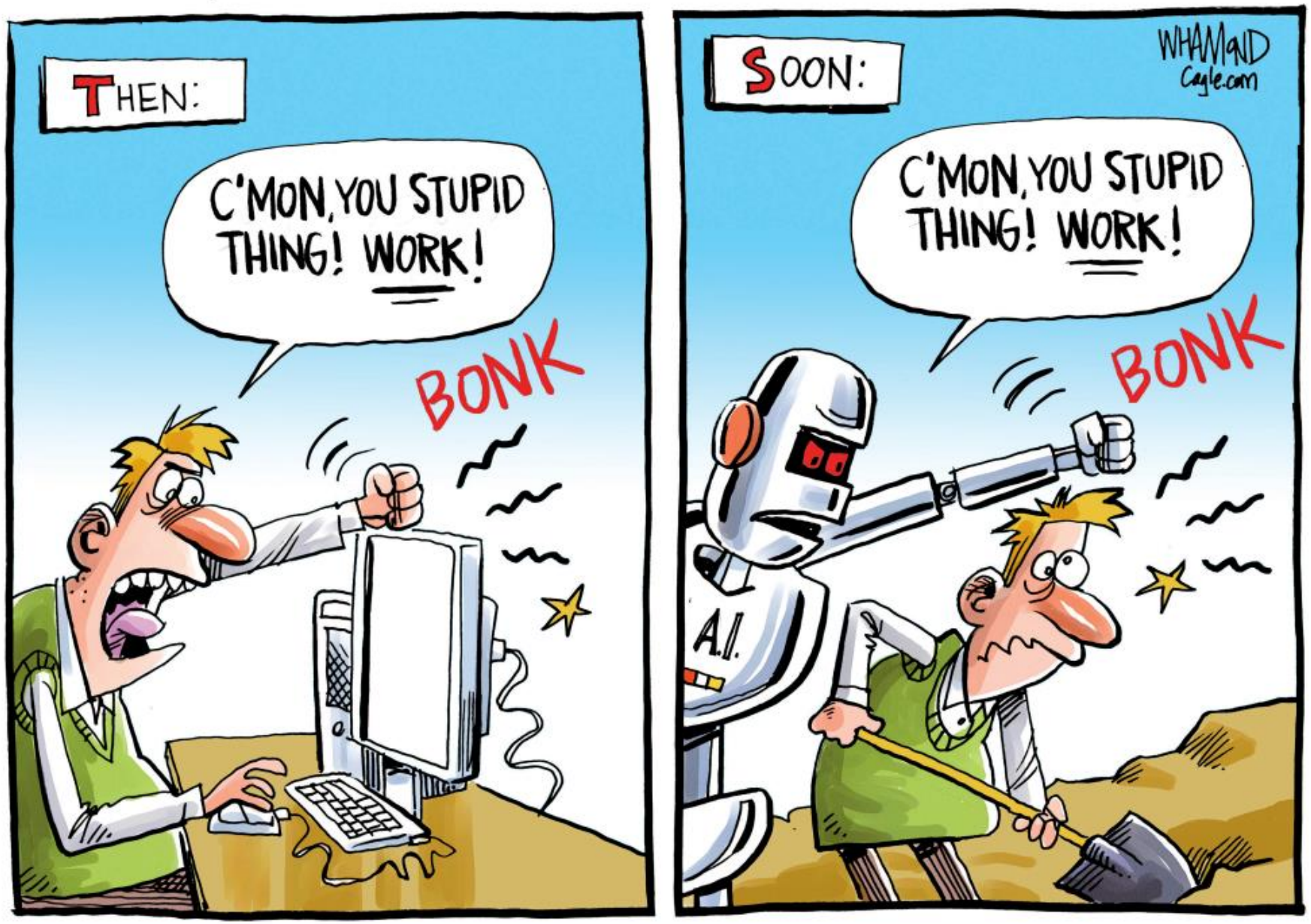
TOMAS FALSAS



POLICÍA NACIONAL

1824 - 2024

TOMAS FALSAS



POLICÍA NACIONAL

TOMAS FALSAS