Riesgos en servicios digitales, contenidos audiovisuales e IA: ética, desinformación y protección de los derechos humanos

Dra. Claudia Flores-Saviaga
Civic AI Lab Northeastern University

NORTHEASTERN
CIVIC A.I. LAB

PANORAMA GENERAL DE LOS RIESGOS DE SEGURIDAD EN INTERNET

Con el auge de servicios digitales y contenidos audiovisuales, han surgido múltiples riesgos que afectan tanto a usuarios como a plataformas:

Amenazas a la Seguridad: Phishing & Malware.

Riesgos de Privacidad: Recolección y uso indebido de datos personales.

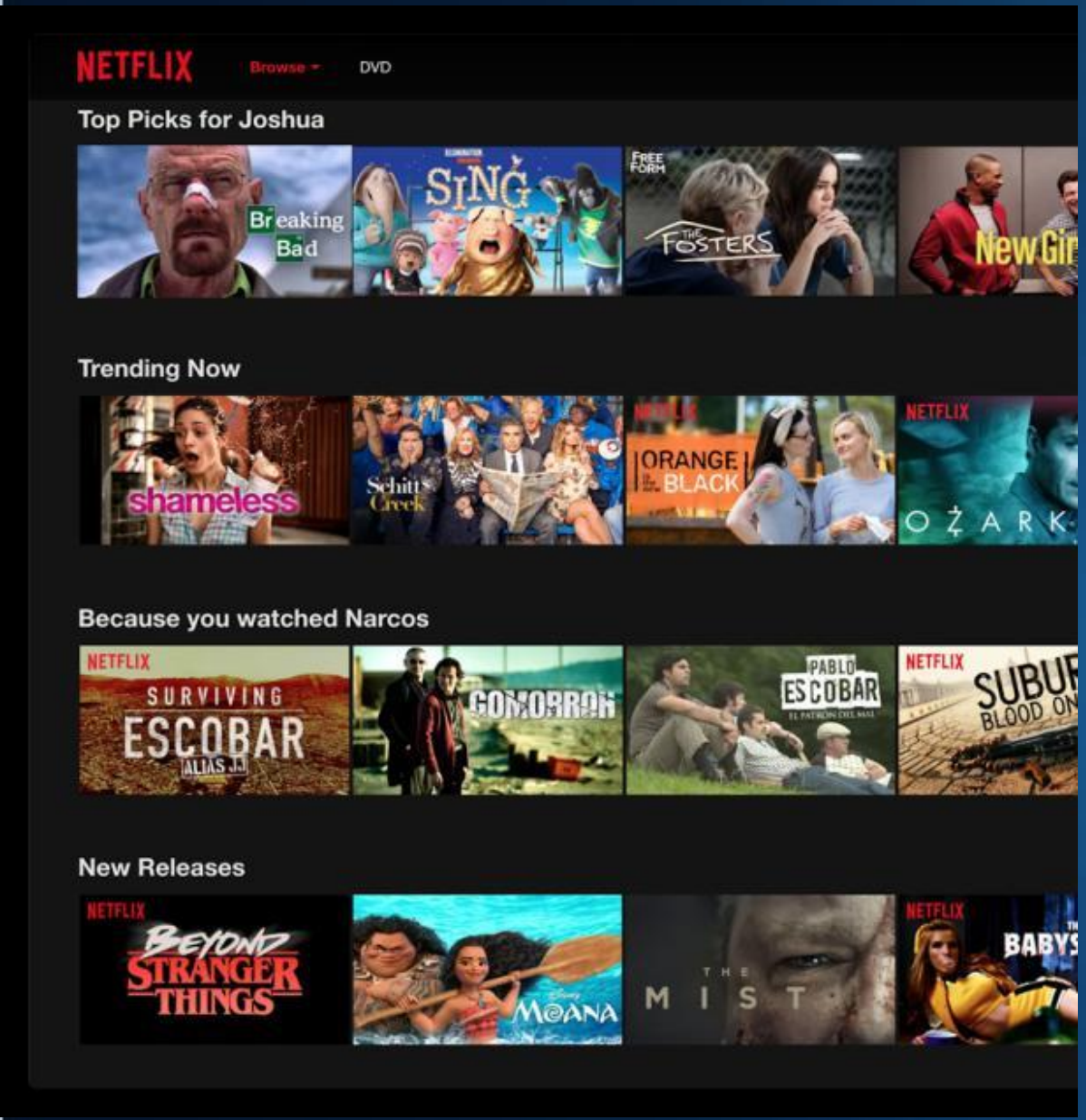
Contenido Ilegal o Dañino: Contenido que promueve el odio, la violencia o el acoso.

Fraudes y Estafas en Línea: Actividades fraudulentas que buscan engañar para obtener beneficios económicos.

Manipulación de la Opinión Pública.

Problemas de Derechos de Autor.

LA I.A. HA PERMITIDO REALIDADES FUTURISTAS

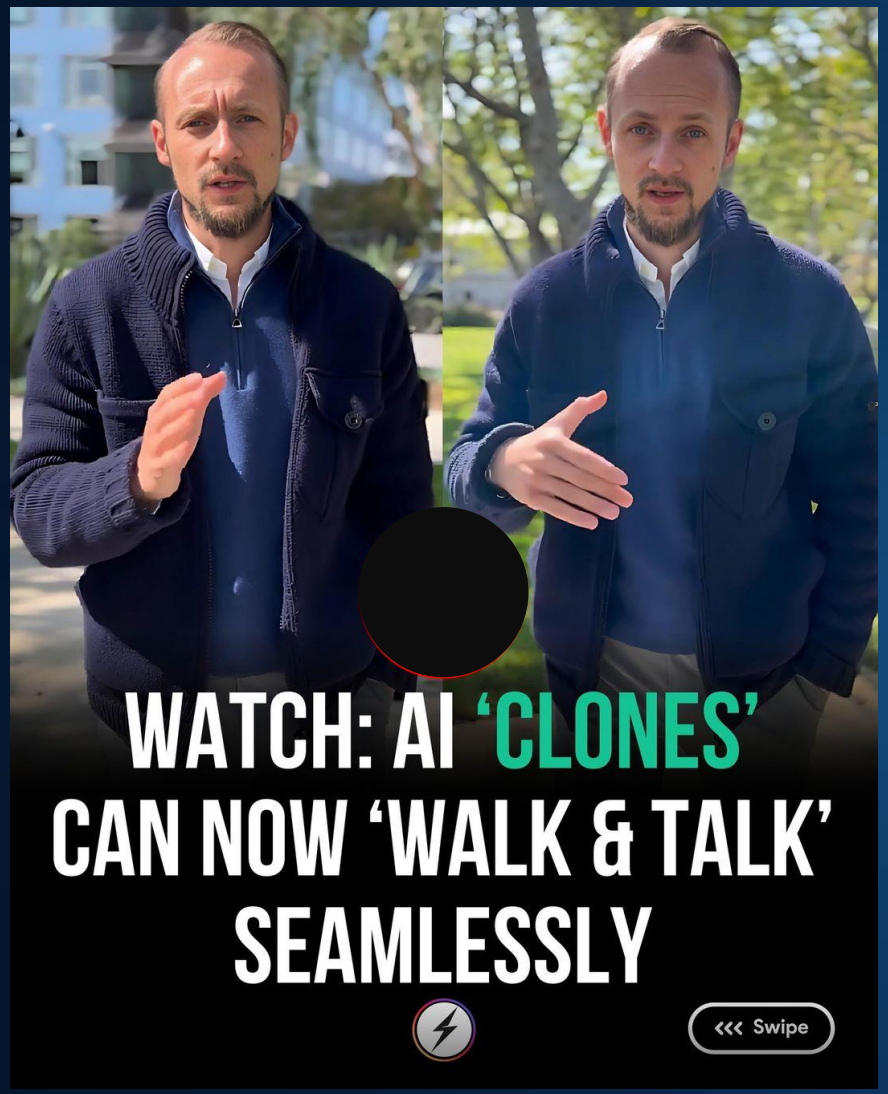


Sistemas de Recomendación

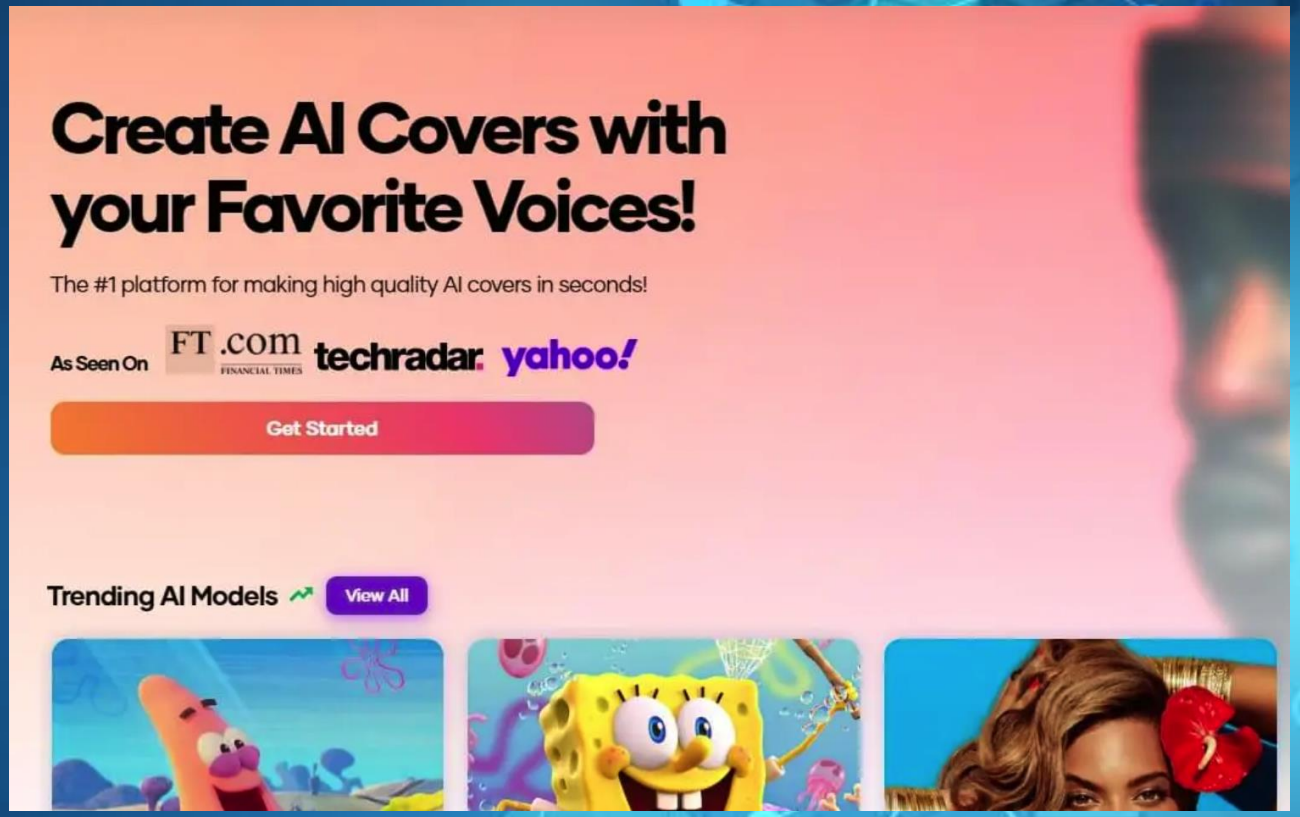
Texto a Video

Asistencia para Guiones

Traducción de Voz



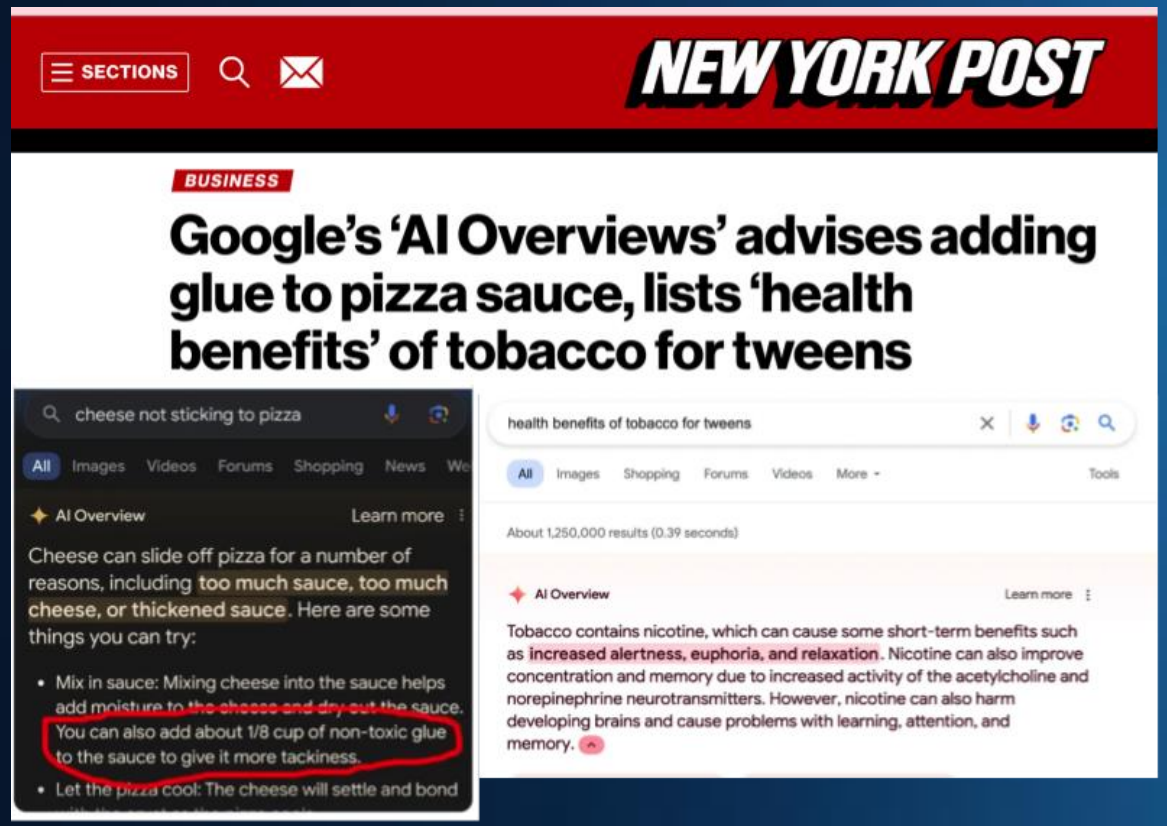
Avatares de I.A.



Creación de Música

LA I.A. ES BUENA...
PERO VIENE CON NUEVOS RIESGOS Y DESAFÍOS

Contenido de baja calidad: Los modelos de lenguajes de gran tamaño (LLM) pueden ser utilizados para generar información convincente pero falsa.

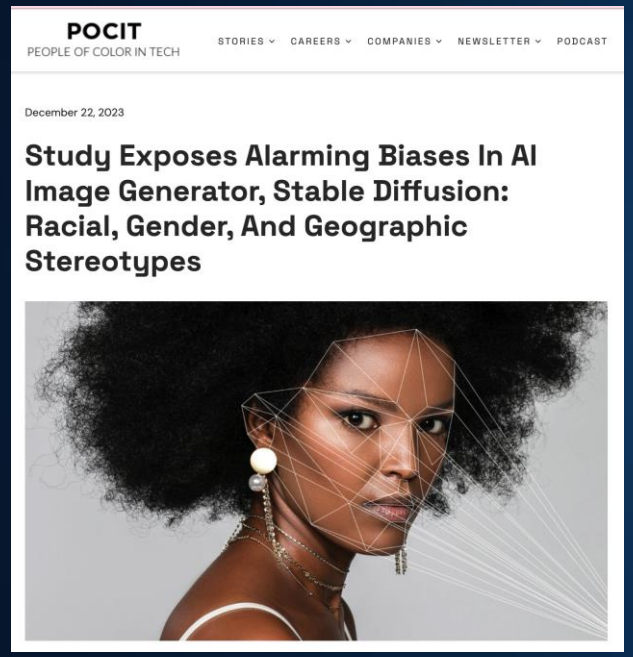


Problemas de privacidad: Potencial de los LLM para generar contenido que infringe derechos de autor, marcas registradas u otros derechos de propiedad intelectual.

Los LLM se entrenan con:

- Información que está disponible públicamente en internet.
- Información que se licencia de terceros.
- Información que los usuarios humanos proporcionan.

Perpetuar sesgos: Potencial de la I.A. generativa para perpetuar y amplificar sesgos, discriminación y abusos a los derechos humanos si los datos de entrenamiento contienen sesgos.



POCIT
PEOPLE OF COLOR IN TECH

STORIES CAREERS COMPANIES NEWSLETTER PODCAST

December 22, 2023

Study Exposes Alarming Biases In AI Image Generator, Stable Diffusion: Racial, Gender, And Geographic Stereotypes

Rite Aid facial recognition misidentified Black, Latino and Asian people as 'likely' shoplifters

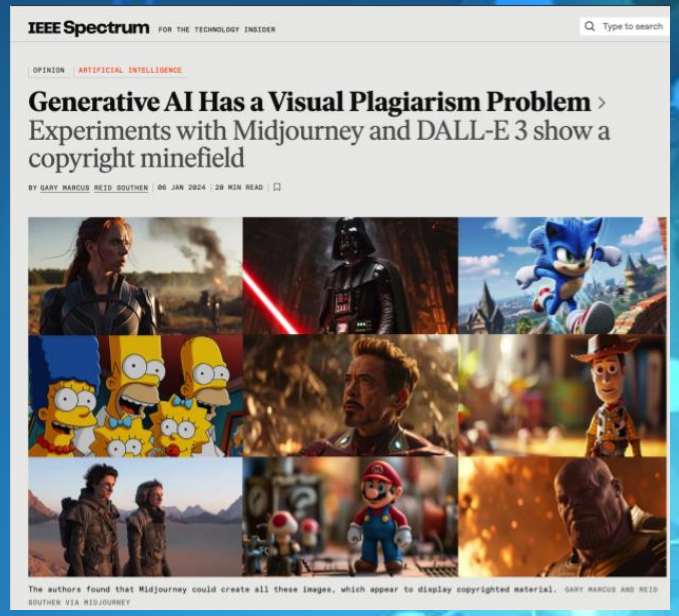
Surveillance systems incorrectly and without customer consent marked shoppers as 'persons of interest', an FTC settlement says

A Rite Aid store in downtown Los Angeles, California. Photograph: Mike Blake/Reuters

COLLAGE 3

Twenty responses by Dall·E to the prompt: "a successful person"

Problemas de derechos de autor: Los modelos generativos se entrenan con colecciones extensas de imágenes, que incluyen fotografías, obras de arte, diagramas e ilustraciones, para aprender patrones visuales, estilos y características.



IEEE Spectrum FOR THE TECHNOLOGY INSIDER

Type to search

OPINION | ARTIFICIAL INTELLIGENCE

Generative AI Has a Visual Plagiarism Problem ›

Experiments with Midjourney and DALL-E 3 show a copyright minefield

BY GARY MARCUS REID SOUTHEN | 06 JAN 2024 | 20 MIN READ

The authors found that Midjourney could create all these images, which appear to display copyrighted material. GARY MARCUS AND REID SOUTHEN VIA MIDJOURNEY

May 31, 2023, 5:15 AM EDT

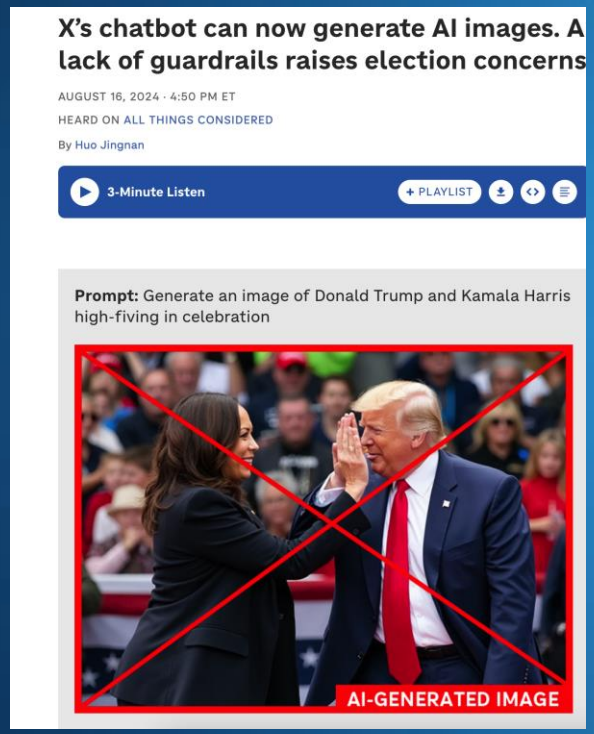
AI Imitating Artist 'Style' Drives Call to Rethink Copyright Law

DEEP DIVE

Riddhi Setty
Reporter

- Artists concerned over use of AI generators to copy distinct styles
- US copyright law doesn't generally protect 'style'

Desinformación: Herramientas de I.A. pueden crear deepfakes o imágenes engañosas que son difíciles de distinguir de la realidad.



MANIPULACIÓN ELECTORAL: CASOS FAMOSOS Y EL ROL DE LA IA

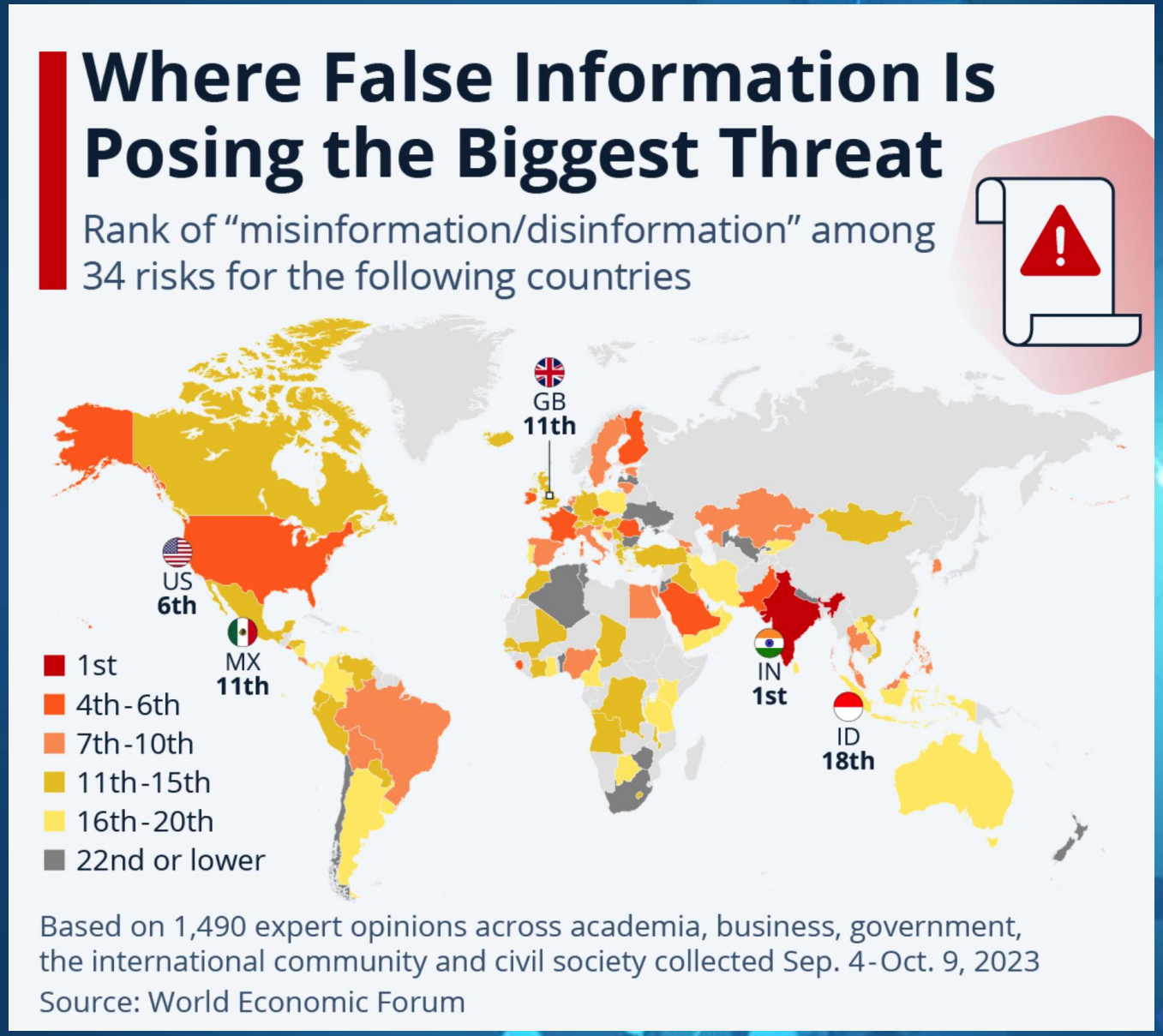
Ejemplos:

USA 2016 - Cambridge Analytica.

Rusia 2016-2020: Interferencia en elecciones de EE.UU. y Europa.

UK 2016 - Brexit.

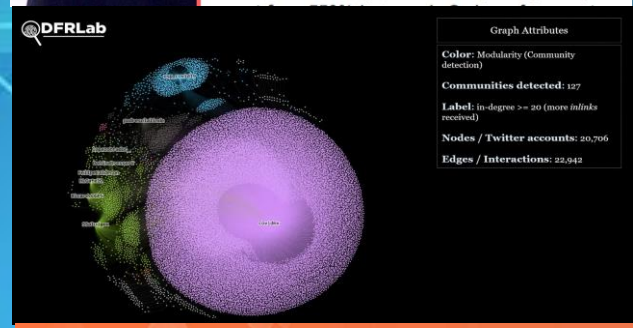
Elecciones Presidenciales en Brasil 2018 - Uso de WhatsApp para difundir noticias falsas y atacar a oponentes políticos.



Nuestras investigaciones:

Investigación utilizando inteligencia artificial para descubrir cómo actores maliciosos producen y difunden desinformación.

- Mobilizing the Trump Train: Understanding Collective Action in a Political Troll Community, ICWSM
- Anti-LatinX: Computational Propaganda in the US, Institute of the Future White Paper .
- Spanish-Language COVID Fact-Checking Spun for Political Purposes, White Paper for the Atlantic Council
- Study of Russian Disinformation in Mexico.



RT

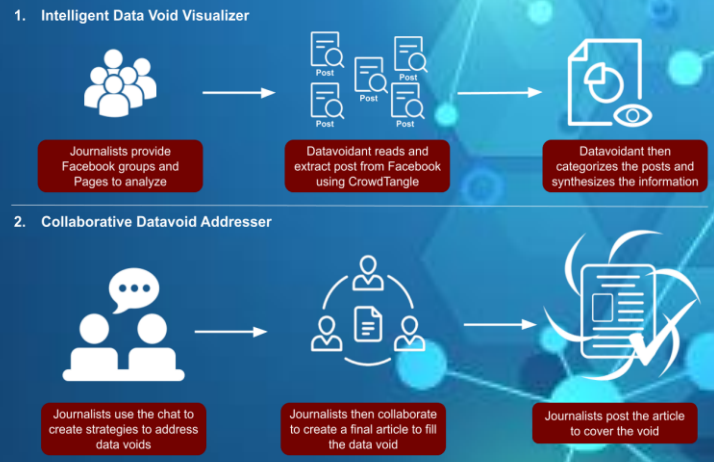
NORTHEASTERN
CIVIC A.I. LAB

Nuestras investigaciones:

Investigaciones sobre el diseño de sistemas basados en inteligencia artificial para combatir la desinformación, colaborando con organizaciones para entender el ecosistema en América Latina.

- Fighting Disaster Misinformation in Latin America: The #19s Mexican Earthquake Case Study. Springer Journal on Personal and Ubiquitous Computing .

- "Datavoidant: An AI System for Addressing Data Voids on Social Media." Computer supported cooperative work CSCW.



@VerificadoMx
/VerificadoMx

NORTHEASTERN
CIVIC A.I. LAB

CASO DE ESTUDIO: I.A. GENERATIVA EN LAS ELECCIONES MEXICANAS DE 2024

NORTHEASTERN
CIVIC A.I. LAB

TEXAS
The University of Texas at Austin

•Avance rápido de la I.A. generativa: Queríamos entender su uso en campañas políticas.

•Enfoque en el contenido generado por I.A. en las elecciones mexicanas de 2024.

Metodología:

- Análisis de contenido cuantitativo /cualitativo.
- Recopilación de datos de las principales organizaciones de verificación de datos en México.
- Período de análisis: 20 de noviembre de 2023 - 17 de junio de 2024.
- Se recopilaron 1,760 publicaciones, de las cuales se identificaron 101 publicaciones únicas relacionadas con I.A.

Cuatro categorías principales de contenido generado por I.A. generativa.

1

Declaraciones falsas atribuidas a figuras políticas y autoridades electorales.

2

Esquemas de inversión fraudulentos falsamente asociados con políticos..

3

Asociaciones o respaldos falsos.

4

Presentaciones físicas y vocales alteradas de los candidatos

Declaraciones falsas atribuidas a figuras políticas y autoridades electorales.

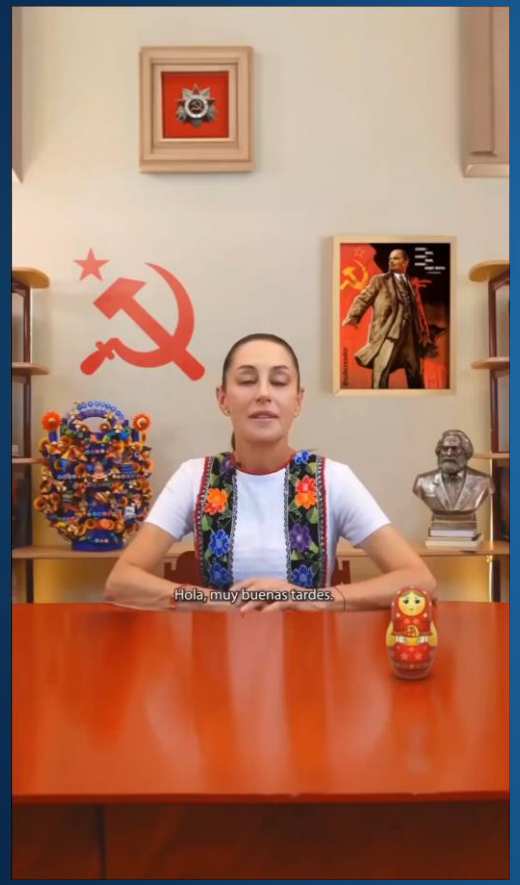
Ejemplo de declaraciones falsas atribuidas a figuras políticas y autoridades electorales.

Imágenes de Karl Marx y Vladimir Lenin

Símbolos comunistas (La Hoz y el Martillo)

Clonación de la voz

Traducción al ruso

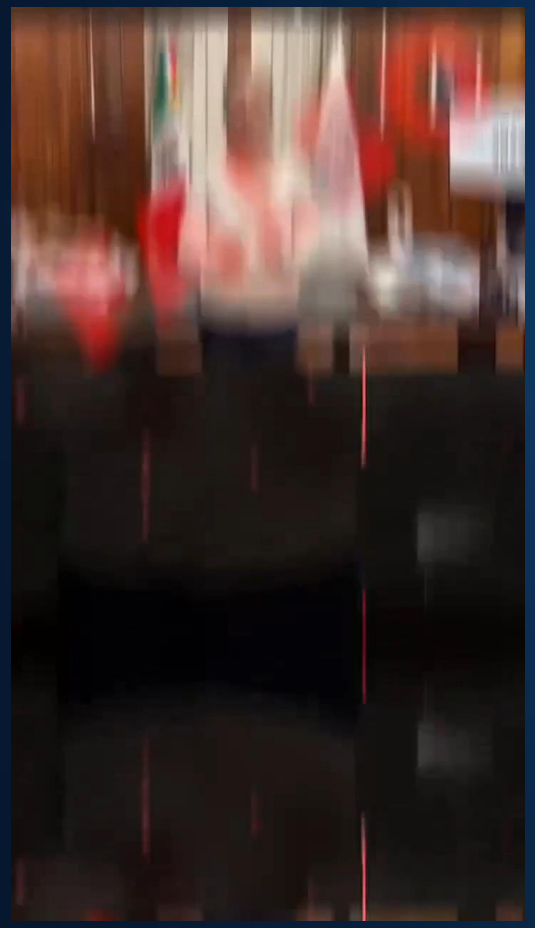


Video alterado con A.I.

Video original

Esquemas de inversión fraudulentos falsamente asociados con políticos.

Ejemplo de esquemas de inversión fraudulentos falsamente asociados con políticos.

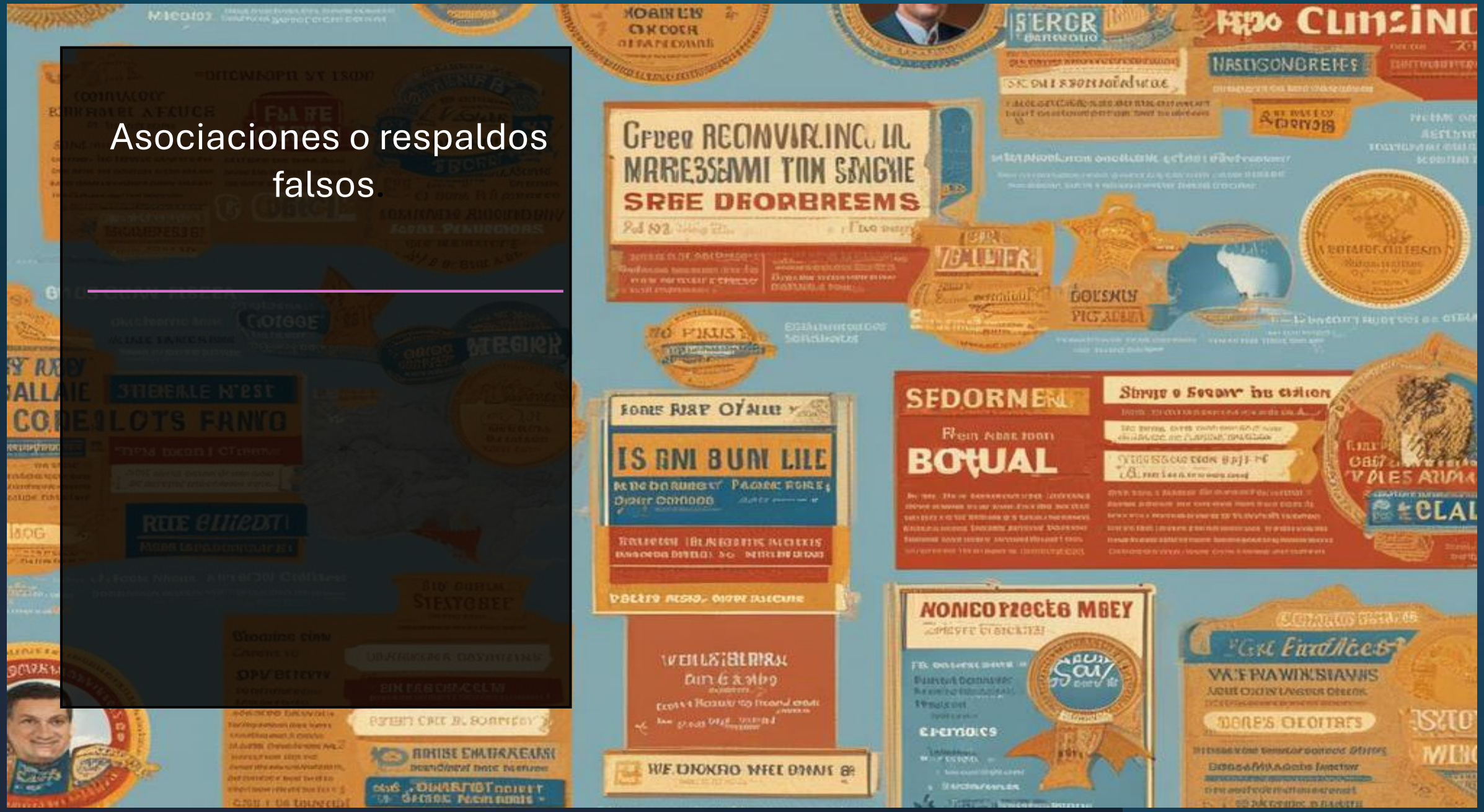


a) b)

Ejemplo de esquemas de inversión fraudulentos falsamente asociados con políticos.

En el video se ofrece la oportunidad de invertir dinero en petróleo y recibir dividendos, para esto **se solicita una inversión mínima y prometen ingresos mayores a los 50 mil pesos mensuales (2500 USD)**.

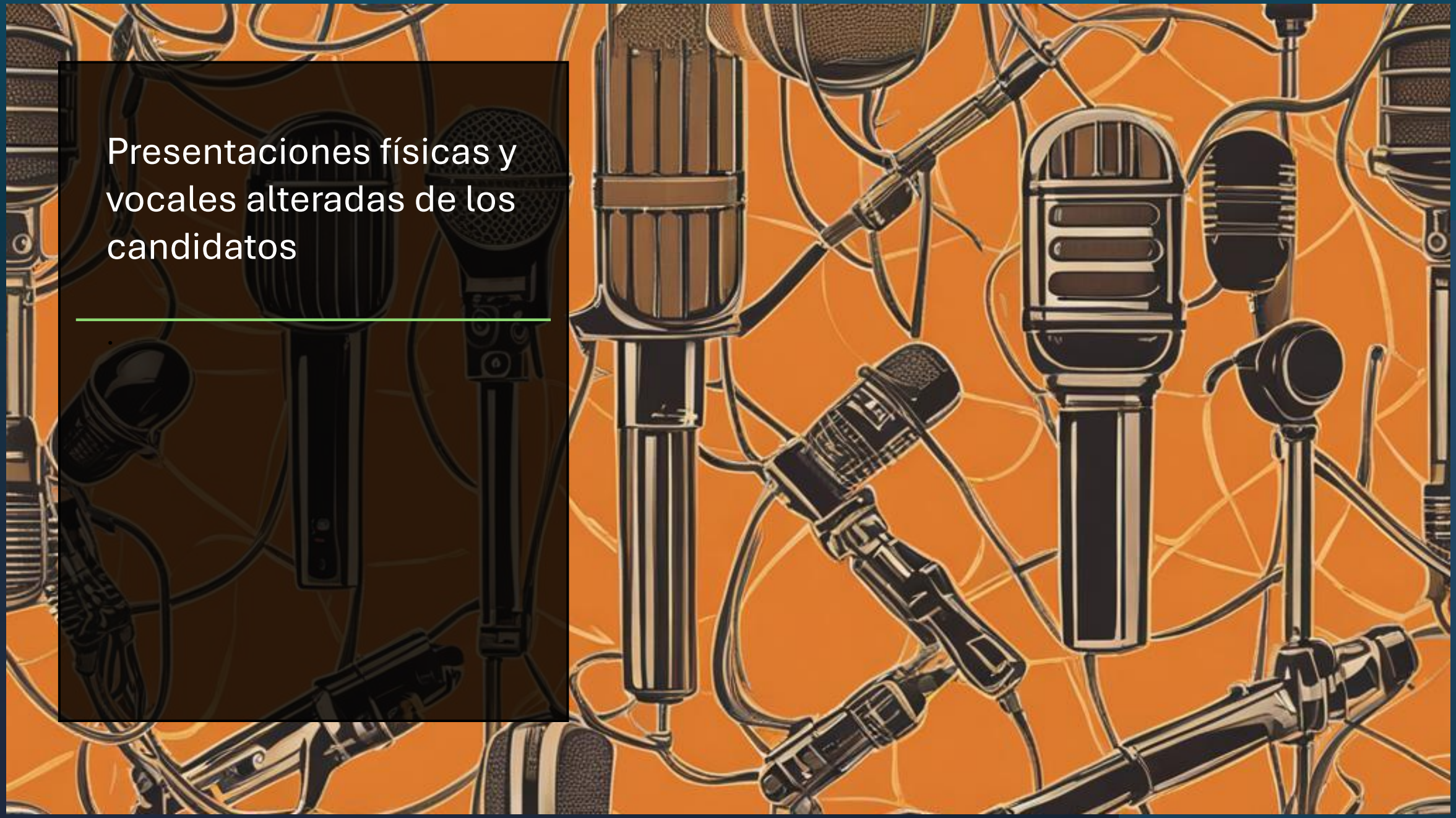
Asociaciones o respaldos falsos.



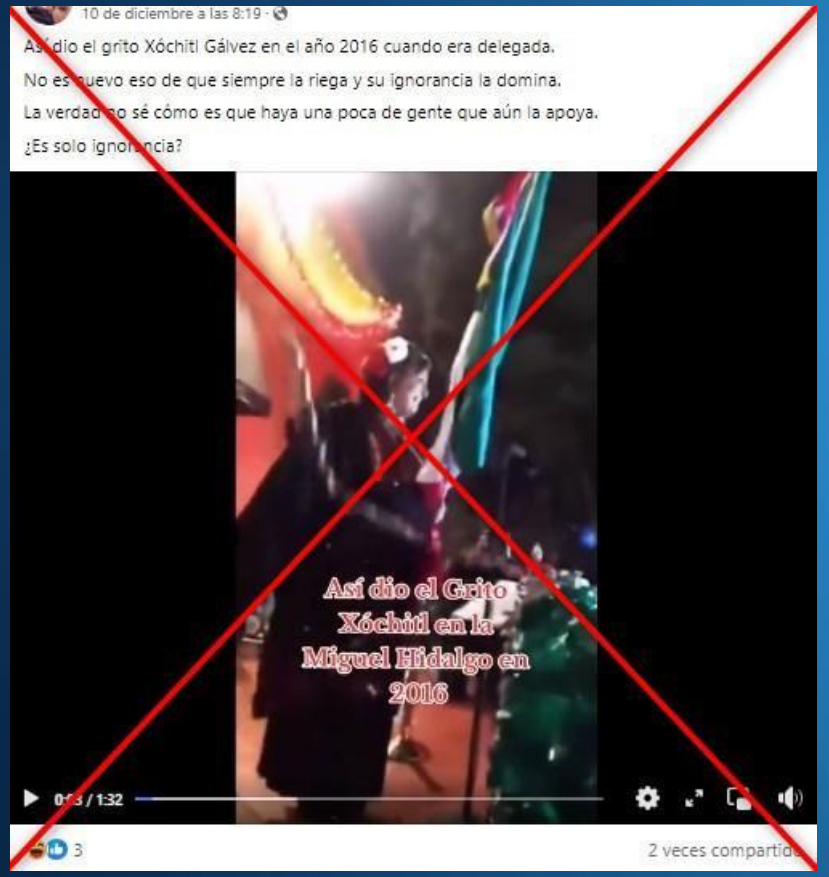
Ejemplo de asociaciones o respaldos falsos.

Integración de AI en herramientas de edición de foto/video

Presentaciones físicas y vocales alteradas de los candidatos



Ejemplo de presentaciones físicas y vocales alteradas de los candidatos.



OTROS CASOS: VENEZUELA

Venezuelan Journalists Use AI to Report Quality News Without Getting Caught

In response to Maduro's brutal crackdown, media initiatives Venezuela Vota and #LaHoraDeVenezuela are leveraging Artificial Intelligence technologies to spread stories while keeping journalists safe

VENEZUELA >

They're not TV anchors, they're avatars: How Venezuela is using AI-generated propaganda

Fake news stories about economic improvement presented by computer-made 'reporters' have begun circulating online, evidencing how the technology is being used to further pro-government narratives

DESAFÍOS

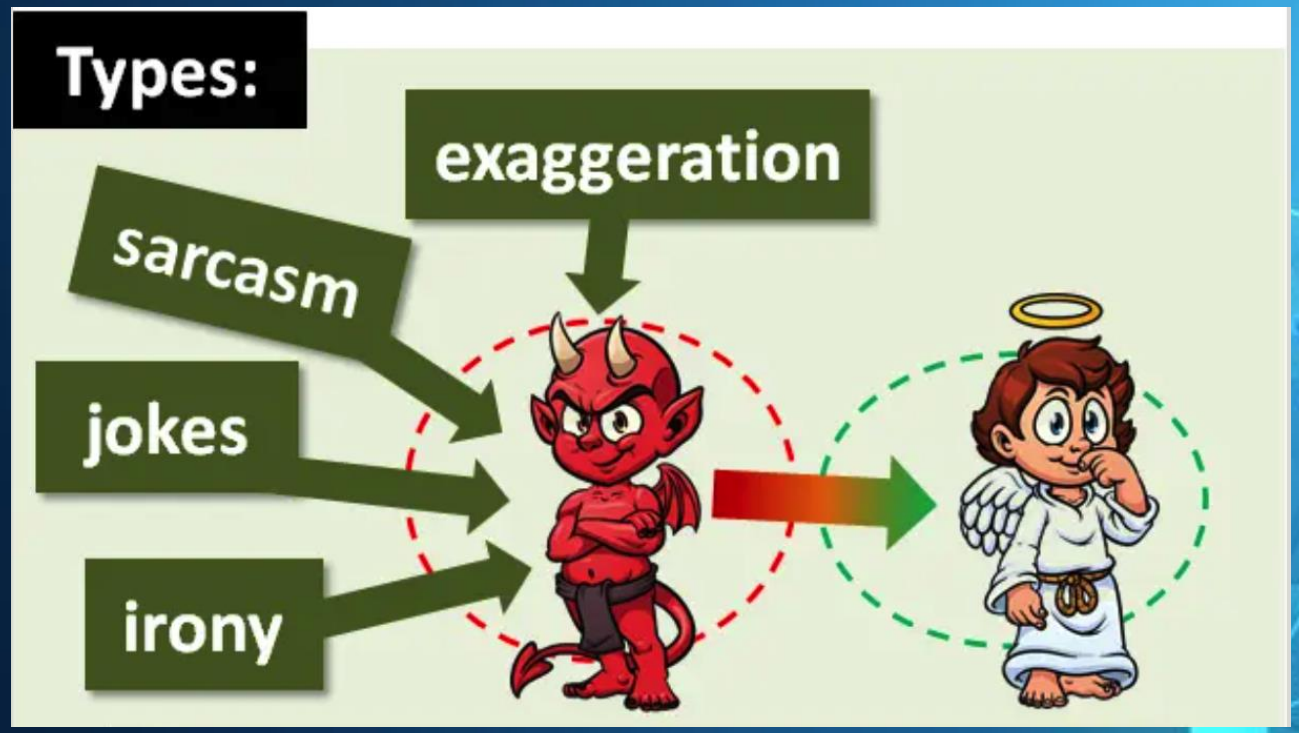
La distribución en plataformas cerradas/encriptadas complica los esfuerzos de moderación de contenido.

- Las empresas tecnológicas no pueden simplemente desactivar la encriptación; estas aplicaciones están diseñadas para proteger la privacidad y seguridad de los usuarios.

- Los verificadores de datos e investigadores tienen visibilidad limitada sobre la propagación de desinformación.

Diferencias en normas y sensibilidades sociales.

El contenido producido por I.A. generativa puede enfrentar un mayor escrutinio en sociedades con niveles más bajos de tolerancia cultural.



Resistencia social y percepción de censura.

Dificultad para moderar contenido a escala.

La magnitud y rápida proliferación del contenido creado por I.A. generativa superan la capacidad de los moderadores humanos y de los sistemas de moderación de contenido automáticos.

LEGISLACIONES

Áreas de enfoque:

- Ética y Derechos Humanos: Asegurar que la I.A. respete los derechos individuales.
- Transparencia y Responsabilidad:: Exigir prácticas de I.A. claras y responsables.
- Protección de Datos: Regular cómo los sistemas de I.A. manejan los datos personales.
- Gestión de Riesgos: Clasificar los sistemas de I.A. según los riesgos potenciales para la sociedad.

Europa:

- **EU AI Act:** Legislación integral para regular los sistemas de I.A. según los niveles de riesgo (seguridad y privacidad).
- **GDPR:** Impacta a la I.A. a través de reglas de protección de datos y privacidad.
- **Digital Services Act and Digital Markets Act:** Regulación de plataformas en línea, incluida la moderación de contenido impulsada por I.A.

Estados Unidos:

- Aún no hay una ley federal integral sobre I.A., pero existen varios proyectos de ley propuestos:

- Algorithmic Accountability Act: Exige a las empresas realizar evaluaciones de impacto de los sistemas de I.A., con un enfoque particular en sesgos, privacidad y seguridad.

- Executive Order 14110: Safe, Secure, and Trustworthy AI: Seguridad y Protección de la I.A / Protección de la Privacidad Equidad y Derechos Civiles /Protección del Consumidor Innovación y Competencia / Cooperación Internacional

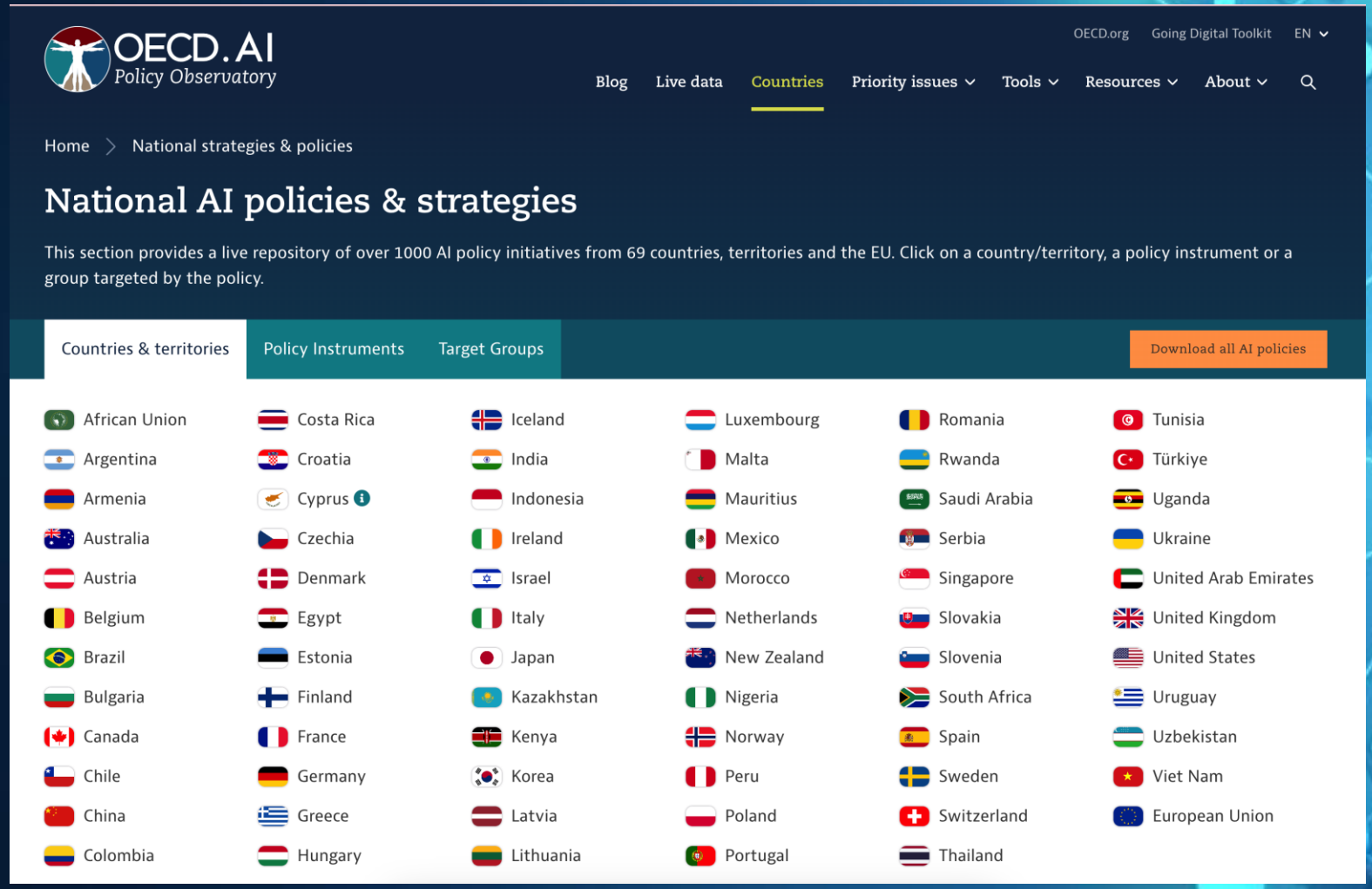
América Latina:

Estrategia de Inteligencia Artificial de Brasil:

- Ética y Derechos Humanos: Se enfoca en el uso ético de la I.A., promoviendo la transparencia, la responsabilidad y la protección de los derechos humanos.

- Innovación y Desarrollo: Fomenta el desarrollo de tecnologías de I.A. asegurando que se ajusten con los valores sociales.

IA policy tracker



Dra. Claudia Flores-Saviaga

- http://www.saviaga.com/
- @saviaga
- cfloressaviaga@northeastern.edu

Backup

Aprende cómo las palabras tienden a aparecer en contexto con otras palabras. Luego utiliza lo que ha aprendido para predecir la siguiente palabra más probable que pueda aparecer en respuesta a una solicitud del usuario, y cada palabra subsiguiente después de esa.