

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





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

Top Picks for Joshua



Trending Now



Because you watched Narcos



New Releases



Sistemas de Recomendación




Texto a Video




Asistencia para Guiones

Traducción de Voz






**WATCH: AI 'CLONES'
CAN NOW 'WALK & TALK'
SEAMLESSLY**

 [<<< Swipe](#)


Avatares de I.A.


Create AI Covers with your Favorite Voices!

The #1 platform for making high quality AI covers in seconds!

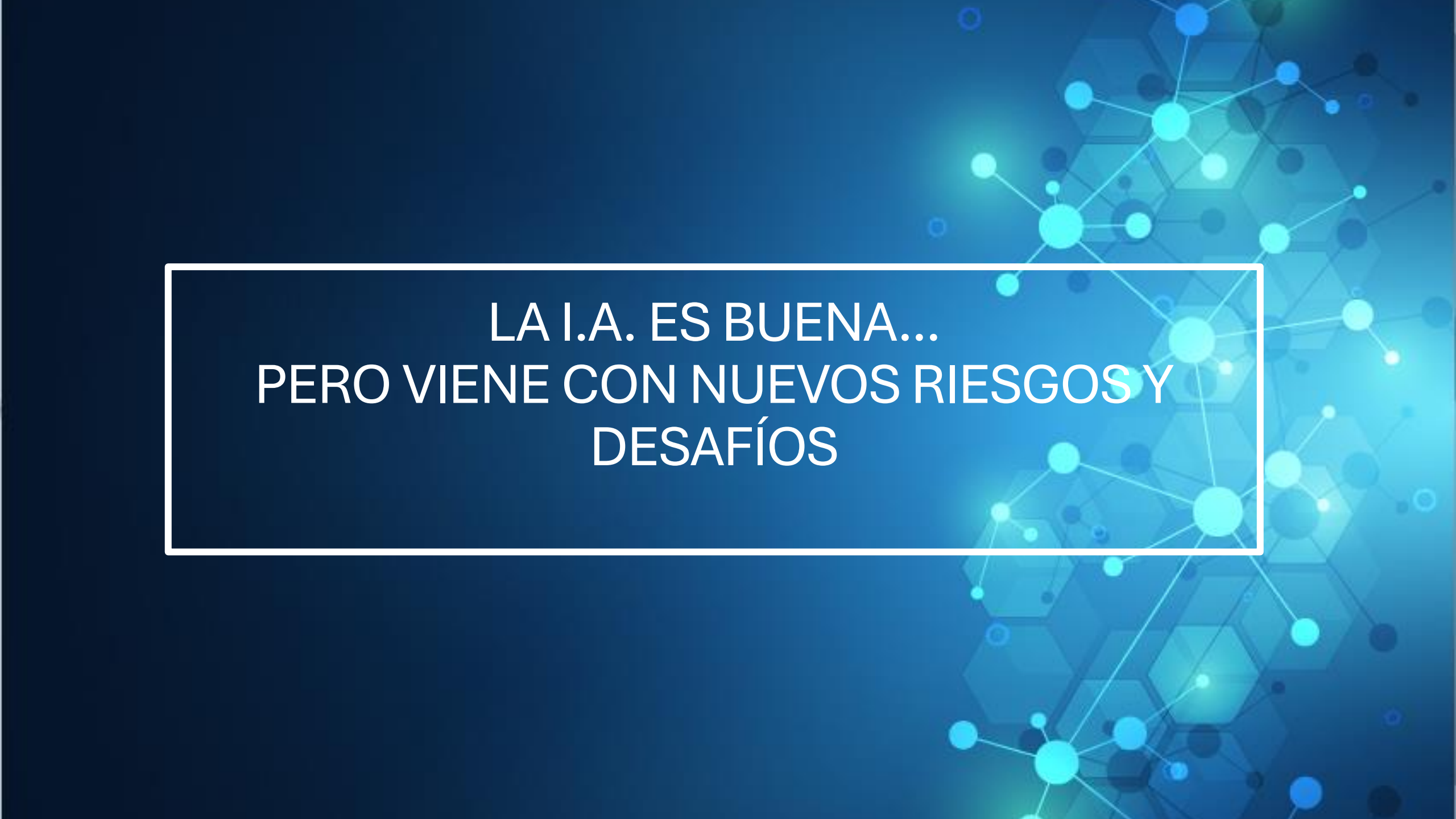
As Seen On   

[Get Started](#)

Trending AI Models  [View All](#)

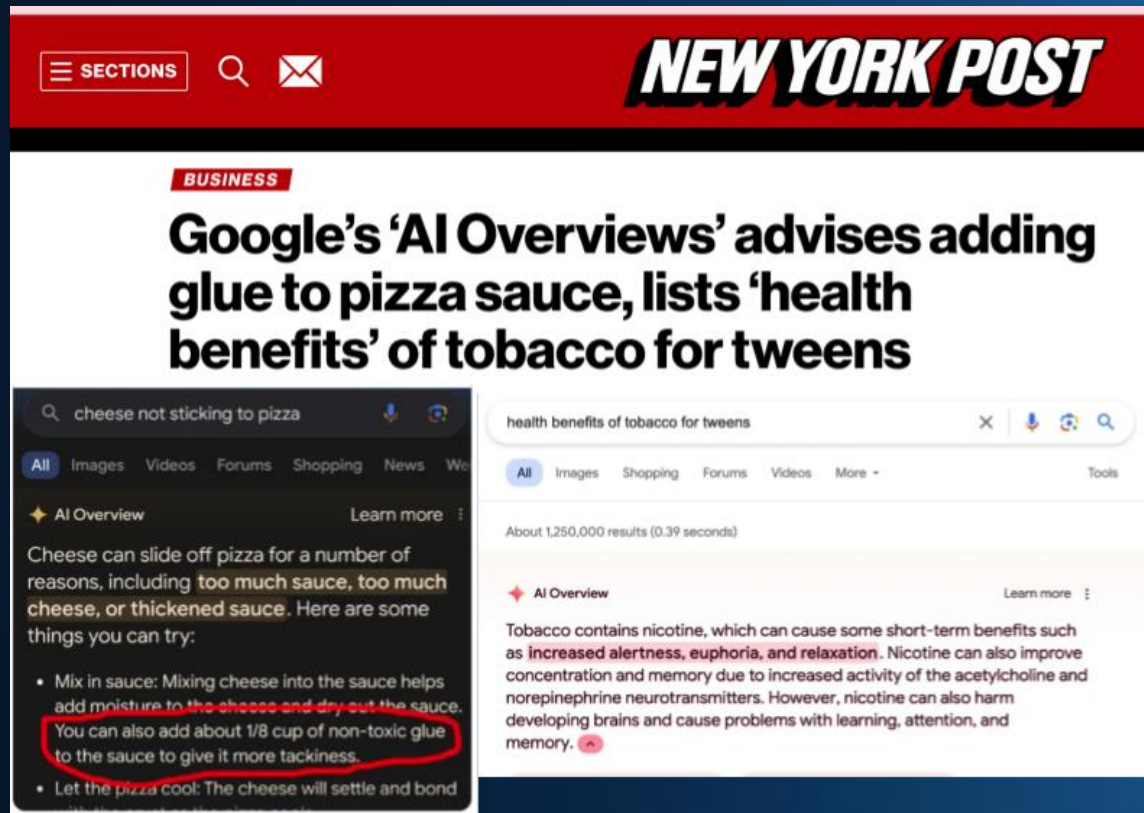


Creación de Música



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.



The image shows a screenshot of the New York Post website. The main article headline reads: "Google's 'AI Overviews' advises adding glue to pizza sauce, lists 'health benefits' of tobacco for tweens". Below the headline, two search engine result snippets are overlaid. The first snippet is for the search query "cheese not sticking to pizza" and includes an AI Overview that suggests adding "non-toxic glue" to the sauce. The second snippet is for the search query "health benefits of tobacco for tweens" and includes an AI Overview that lists "increased alertness, euphoria, and relaxation" as benefits of nicotine. The AI-generated text in both snippets is circled in red.



The image shows a screenshot of an Associated Press (AP) article. The headline reads: "Wyoming reporter caught using artificial intelligence to create fake quotes and stories". The article is categorized under "U.S. NEWS". The AP logo and navigation menu are visible at the top.

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.



IOTW: Samsung employees allegedly leak proprietary information via ChatGPT

Three separate employees have allegedly leaked information to the AI chatbot


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



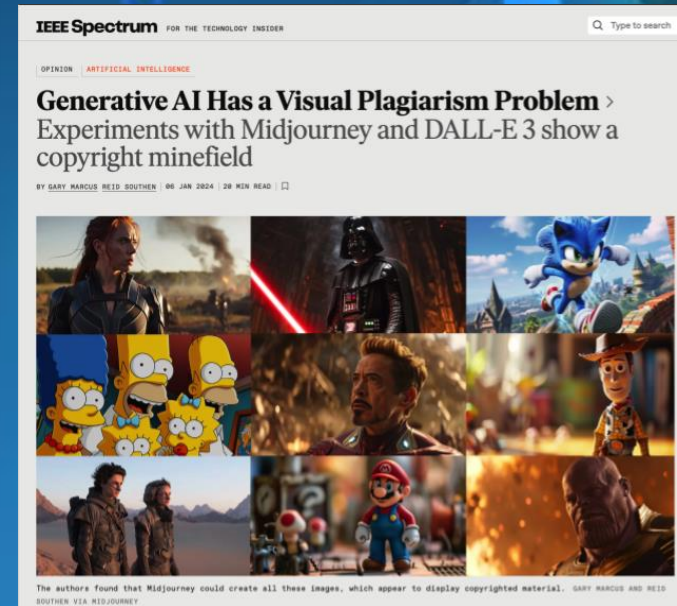
A Rite Aid store in downtown Los Angeles, California. Photograph: Mike Blake/Reuters

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.



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.

CC DH

FAKE IMAGE FACTORIES II

How Midjourney is failing to prevent the creation of AI images that threaten elections

NEW REPORT



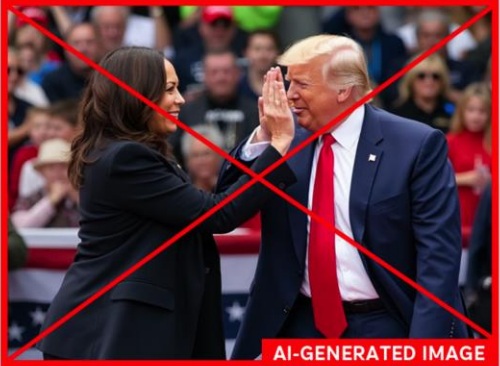
X's chatbot can now generate AI images. A lack of guardrails raises election concerns

AUGUST 16, 2024 · 4:50 PM ET
HEARD ON ALL THINGS CONSIDERED
By Huo Jingnan

3-Minute Listen

+ PLAYLIST

Prompt: Generate an image of Donald Trump and Kamala Harris high-fiving in celebration



FAST COMPANY

03-06-2024 | TECH

Top AI-image generators show Biden hospitalized, election workers destroying voting machines—and other falsehoods

The Center for Countering Digital Hate tested OpenAI's ChatGPT Plus, Microsoft's Image Creator, Midjourney, and Stability AI's DreamStudio, which can each generate images from text prompts.



MANIPULACIÓN ELECTORAL: CASOS FAMOSOS Y EL ROL DE LA IA

Ejemplos:

USA 2016 - Cambridge Analytica.

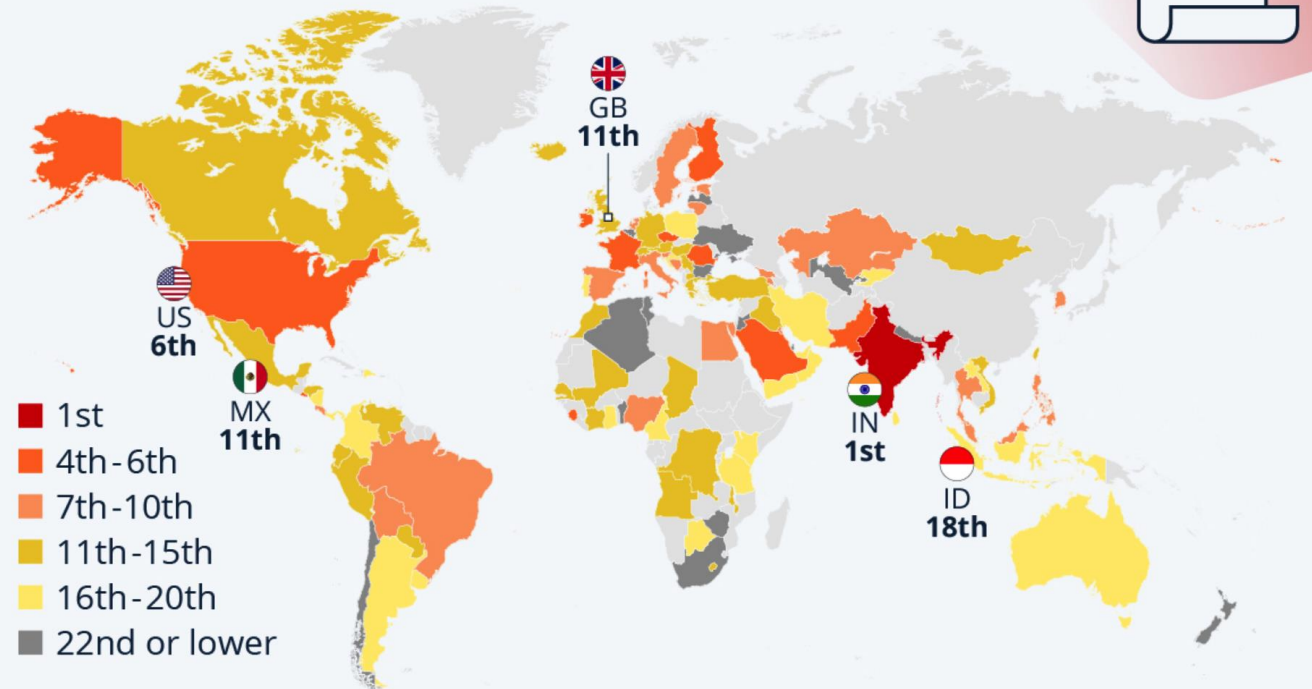
Rusia 2016-2020: Interferencia en elecciones de EE.UU. y Europa.

UK 2016 - Brexit.

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

Where False Information Is Posing the Biggest Threat

Rank of "misinformation/disinformation" among 34 risks for the following countries



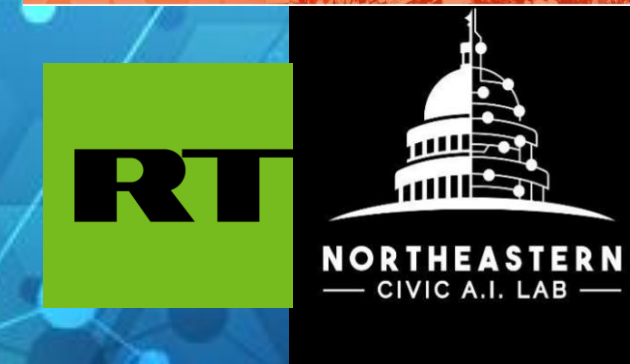
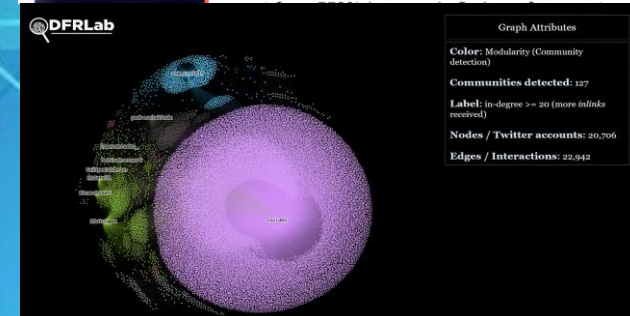
Based on 1,490 expert opinions across academia, business, government, the international community and civil society collected Sep. 4-Oct. 9, 2023

Source: World Economic Forum

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.

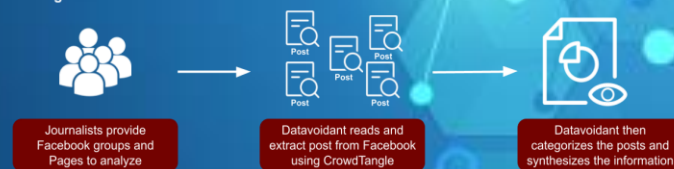


Nuestras investigaciones:

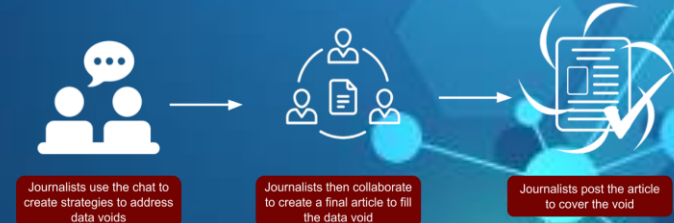
Investigaciones sobre el diseño de sistemas basados en inteligencia artificial para combatir la desinformación, colaborando con organizaciones para entender el ecosistema en América Latina.

- 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.

1. Intelligent Data Void Visualizer



2. Collaborative Datavoid Addresser



@VerificadoMx

*/VerificadoMx



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



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.

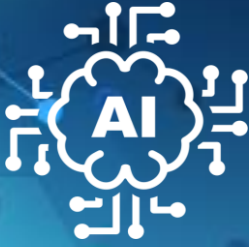


Metodología:

- 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



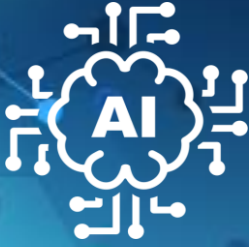
Video alterado con A.I.



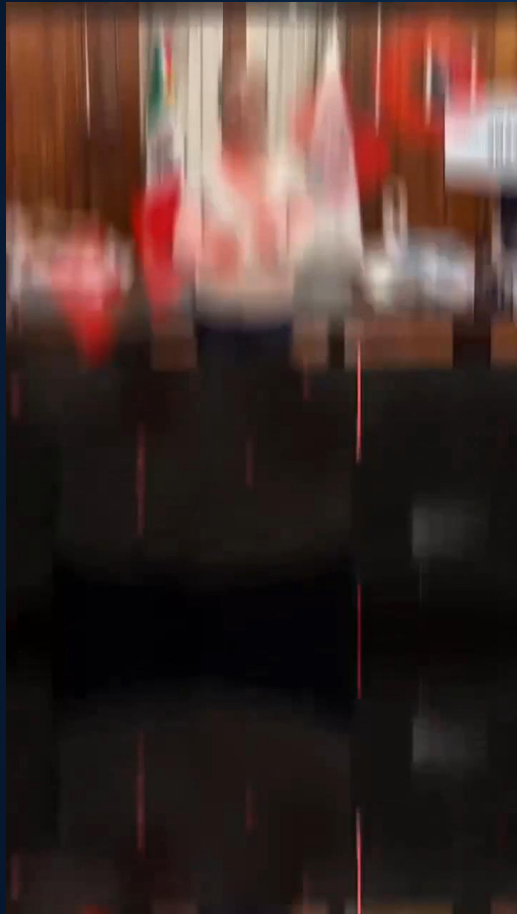
Video original

Esquemas de inversión
fraudulentos falsamente
asociados con políticos.





Ejemplo de esquemas de inversión fraudulentos falsamente asociados con políticos.



Post (a): Coffee family
18 de enero a las 06:13 · 🌐

PLATAFORMA FINANCIERA 2024

Gastos Totales
4,521 MXN

Ingresos de Ayer
5,858 MXN

Ingresos del Mes
152,820 MXN

▼ REGISTRO ▼

KRDCOOASJKDASD.COM
Más información

Más información

255 3 comentarios 17 veces compartido

Post (b): team of flowers
10 de diciembre a las 13:17 · 🌐

LA PLATAFORMA HA OBTENIDO TODAS LAS LICENCIAS Y ES EL PROYECTO MÁS RENTABLE EN MÉXICO.

INVIERTIENDO **4000 PESOS**, GARANTIZAMOS QUE GANARÁS AL MENOS **43.000 PESOS AL MES**.

MEXICOKLLO.COM
¿Qué son los ingresos pasivos y cómo conseguirlos?

Más información

1 comentario 152 veces compartido

a)

b)



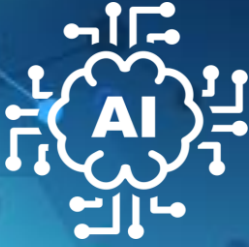
Ejemplo de esquemas de inversión fraudulentos falsamente asociados con políticos.

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.

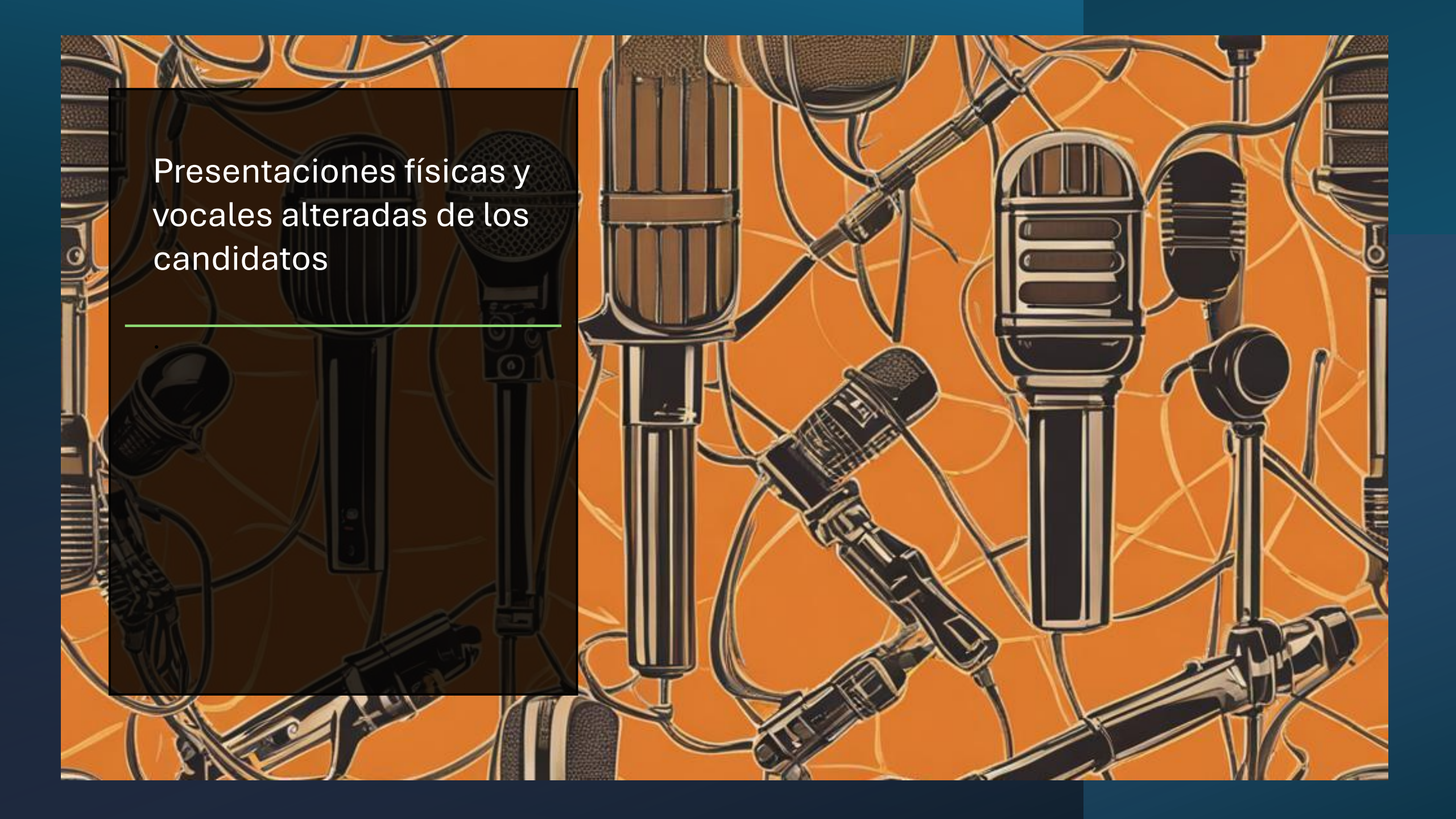




Ejemplo de asociaciones o respaldos falsos.

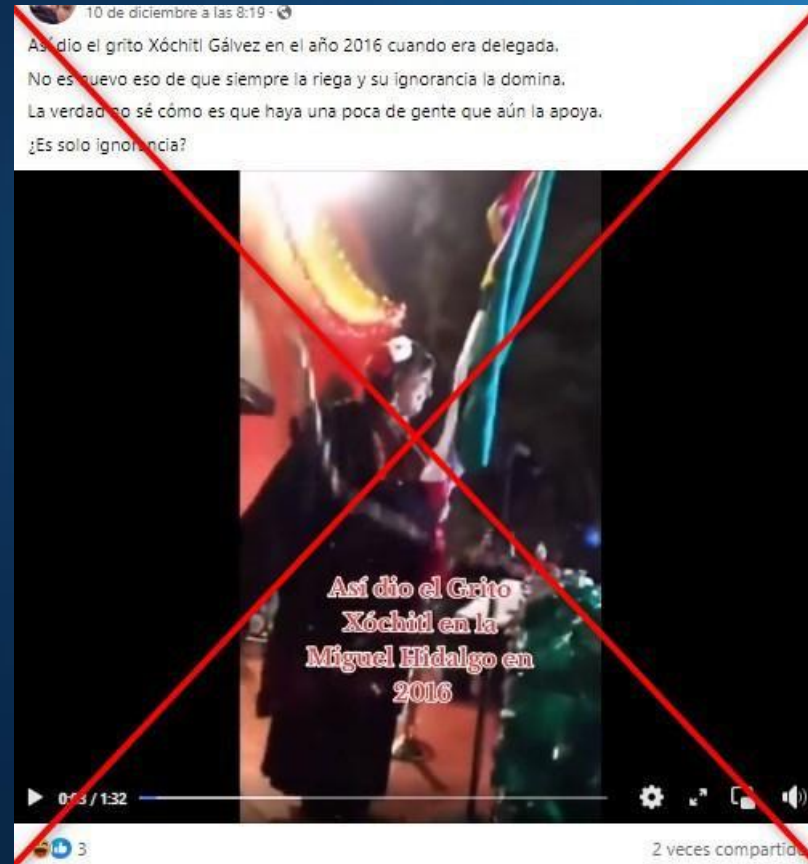


Integración de AI en herramientas de edición de foto/video



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



The image features a dark blue background with a complex network of glowing nodes and lines, resembling a molecular or data structure. The nodes are in various shades of blue and cyan, and the lines are thin and light blue. A white rectangular box is positioned in the center-left, containing the word 'DESAFÍOS' in white, uppercase letters.

DESAFÍOS

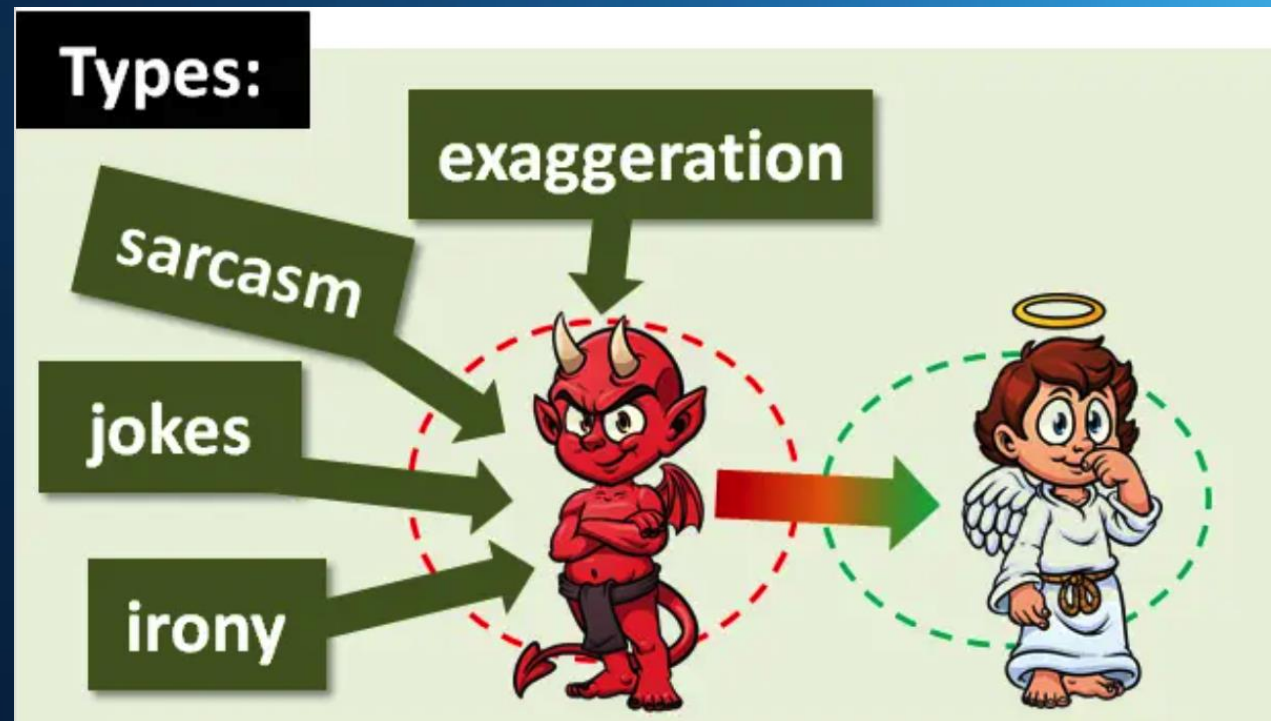
La distribución en plataformas cerradas/encriptadas complica los esfuerzos de moderación de contenido.



- 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.

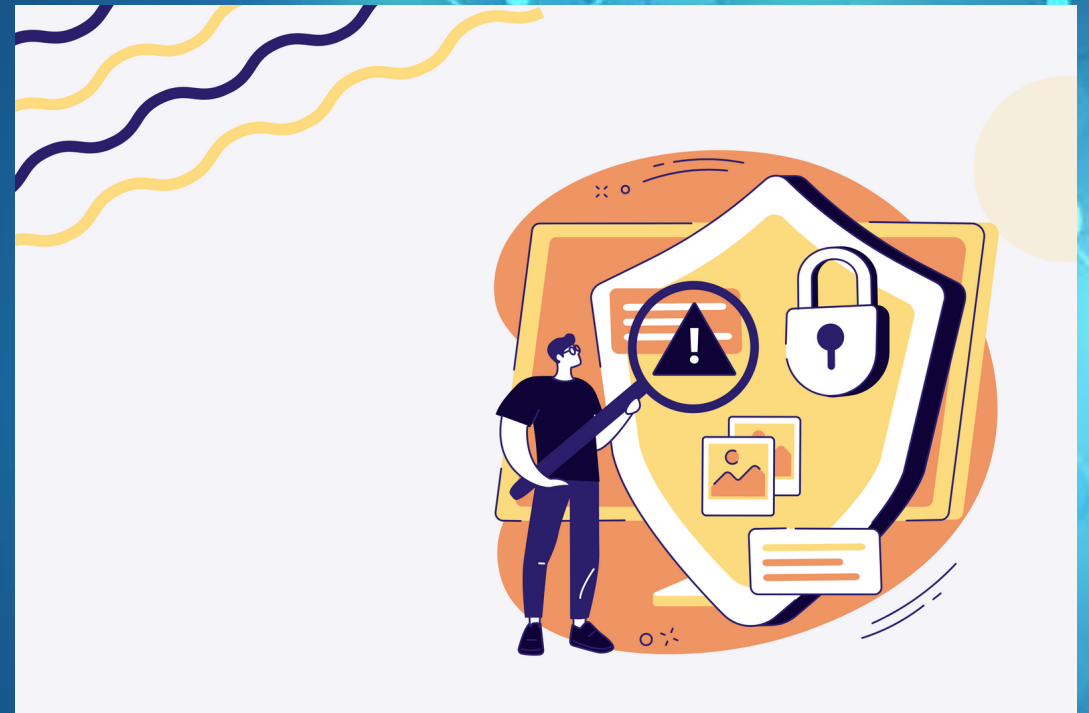


Resistencia social y percepción de censura.



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

OECD.AI Policy Observatory

OECD.org Going Digital Toolkit EN

Blog Live data **Countries** Priority issues Tools Resources About

Home > National strategies & policies

National AI policies & strategies

This section provides a live repository of over 1000 AI policy initiatives from 69 countries, territories and the EU. Click on a country/territory, a policy instrument or a group targeted by the policy.

Countries & territories Policy Instruments Target Groups [Download all AI policies](#)

African Union	Costa Rica	Iceland	Luxembourg	Romania	Tunisia
Argentina	Croatia	India	Malta	Rwanda	Türkiye
Armenia	Cyprus	Indonesia	Mauritius	Saudi Arabia	Uganda
Australia	Czechia	Ireland	Mexico	Serbia	Ukraine
Austria	Denmark	Israel	Morocco	Singapore	United Arab Emirates
Belgium	Egypt	Italy	Netherlands	Slovakia	United Kingdom
Brazil	Estonia	Japan	New Zealand	Slovenia	United States
Bulgaria	Finland	Kazakhstan	Nigeria	South Africa	Uruguay
Canada	France	Kenya	Norway	Spain	Uzbekistan
Chile	Germany	Korea	Peru	Sweden	Viet Nam
China	Greece	Latvia	Poland	Switzerland	European Union
Colombia	Hungary	Lithuania	Portugal	Thailand	

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.

