EFCSN

Cómo detectar y entender contenido generado por IA o alterado digitalmente

Los modelos de IA accesibles y fáciles de usar pueden ayudar a las personas a aprender y a crear contenido, pero también pueden **amplificar los riesgos que supone la desinformación** para las sociedades abiertas y el discurso democrático. Es importante evitar que nuestros espacios de información compartidos se llenen de desinformación generada con IA o digitalmente alterada.

Parte de la solución son nuevas tecnologías como las iniciativas de autenticidad del contenido y *software* de detección. **Pero las soluciones tecnológicas están lejos de ser perfectas y necesitamos el trabajo de verificadores independientes para ofrecer a la sociedad un conjunto compartido de hechos verificados.**

Aquí tienes un resumen rápido de lo que hacen los profesionales del *fact-checking* independientes para identificar y desmentir la desinformación generada con IA, y lo que puedes aprender de ellos.

Aumenta el contenido generado con IA

Hoy en día, la desinformación generada con IA constituye un pequeño porcentaje de todas las afirmaciones investigadas por *fact-checkers* profesionales e independientes. El contenido alterado digitalmente es más prevalente en el trabajo de los verificadores.

Pero: en una encuesta interna a miembros de EFCSN, **la mayoría de los verificadores coincidieron en que el contenido generado por IA y alterado digitalmente aumentará su peso en el futuro**. Y ejemplos recientes en el contexto de las elecciones europeas respaldan este pronóstico.

IMPORTANTE: *Alterado digitalmente* se refiere a cualquier forma de contenido que ha sido alterado de manera significativa para manipular o cambiar el mensaje que originalmente transmitía, incluyendo ediciones con herramientas de IA. Esto no incluye ediciones para darle mayor claridad o calidad.

Generado con IA se refiere a cualquier forma de contenido que ha sido creado por un sistema de inteligencia artificial.

Información facilitada por AI@EUElections, financiado por Meta y gestionado por la **European Fact-Checking Standards Network** en asociación con sus miembros y expertos externos.

La tecnología avanza rápido, pero no podemos depender solo de ella

Los expertos en IA y los verificadores profesionales coinciden: **las herramientas de detección de IA no son suficientes por sí solas para detectar o desmentir contenido generado por IA o alterado digitalmente.**

Antes de usar un detector, los expertos recomiendan familiarizarse con los generadores y detectores de contenido de IA. Entendiendo cómo se entrenan los modelos y un poco de estadística, los *fact-checkers* pueden empezar a reconocer las fortalezas y debilidades de una herramienta, y su probabilidad de éxito. **Aun así, las herramientas pueden ser un buen punto de partida.**

Las iniciativas de autenticidad de contenido, como las especificaciones de C2PA, pueden ayudar a certificar la fuente y el historial de un contenido, pero la marca de agua y la verificación no son infalibles.

Cómo afecta a las personas la desinformación generada con IA

"Cada vez que reaccionas, por así decirlo, con instinto, se salta tu reflexión".

\- Christine Dugoin*

PSICOLOGÍA: Las operaciones de influencia a menudo están diseñadas para aprovechar los sesgos psicológicos.
Entender los tuyos propios, y los de tu audiencia, puede ayudar a contrarrestar la desinformación.

OBJETIVOS: ¿Por qué un actor malintencionado podría recurrir a la IA para crear o difundir desinformación? ¿Cuál es el impacto que se busca en el mundo real?

- ➔ ¿Expandir su alcance a otro país o comunidad?
- ➔ ¿Evitar la detección o abrumar a los *fact-checkers* generando muchas variantes de afirmaciones similares?
- ➔ ¿Influir en los pensamientos generando credibilidad a través de redes sociales con cuentas falsas?

** Christine Dugoin es investigadora especializada en influencia informativa de La Sorbona.*

Desmentir requiere un enfoque multifacético y una comprensión matizada

Si las herramientas de detección no funcionan, ¿qué funciona? Es importante entender el contexto de una afirmación tanto como su contenido. Los *fact-checkers* son expertos en investigación. Aquí tienes algunos consejos.

> *"Las herramientas de detección nunca funcionarán al 100% - no espero que lo hagan".*
>
> - Henk van Ess**

EVALÚA LA FUENTE: ¿Puedes confirmar su identidad? ¿De qué hablan y qué comparten? ¿Quién interactúa con su contenido? ¿Qué efecto podría tener este contenido en los lectores?

FIJA LA CREDIBILIDAD: Verifica la información de forma independiente con fuentes creíbles, como expertos con práctica en el campo. ¿Lo que se muestra tiene sentido según tu conocimiento?

Usa técnicas **FORENSES** para complementar la investigación periodística tradicional y la investigación documental. Algunas técnicas incluyen: extracción de datos, geolocalización, reconocimiento biométrico, análisis de patrones y más.

APRENDE Y ADÁPTATE: Los creadores de desinformación generada con IA se adaptan constantemente. Ajusta tu enfoque a este escenario cambiante.

COMPARTE TU TRABAJO

Junto con una afirmación desmentida, los expertos recomiendan proporcionar un análisis transparente y enlaces a fuentes. Esto puede ayudar a los lectores a seguir la investigación y a entender los matices de una narrativa. En algunos casos, la investigación es más importante que si el contenido está escrito con IA.

** *Henk van Ess es experto en OSINT y técnicas de* fact-checking.

Guía rápida: qué hacer, qué no hacer y algunas pistas

Las siguientes son pistas que podrían indicar que un contenido ha sido generado con IA o alterado digitalmente. Junto con los otros consejos de esta guía (contexto, técnicas de investigación y herramientas de detección), pueden ayudarte a entender la verdad detrás de lo que ves.

Texto

- A menudo (pero no siempre) tiene **mejor gramática** que la de un humano.
- Probablemente use **lenguaje demasiado formal o estructurado**, especialmente en un contexto de redes sociales.
- **Exceso de adverbios o adjetivos.**
- Falta de emoción, humor, sarcasmo y expresiones idiomáticas humanas.
- Puede **carecer de detalles específicos** (nombres, fechas, lugares) o ideas originales.
- Lo más importante: ¿los hechos mencionados son correctos?

Vídeo

- No uses un detector de imágenes generadas con IA en fotogramas de vídeo.
- Observa las **expresiones faciales y los movimientos,** como parpadeos o si el movimiento de la boca coincide con el audio.
- Puede contar con **transiciones o cortes abruptos.**

Audio

- **Compara el audio dudoso con una muestra auténtica** usando herramientas que puedan detectar diferencias en patrones de habla, respiración, entonación...
- Al usar detectores, evita muestras de audio de baja calidad y con ruido de fondo.
- Puede contener **patrones de habla antinaturales o mecánicos,** falta de pausas o de respiración natural.

Imagen

- Busca áreas con **detalles antinaturales:** piel perfecta, fondos borrosos, belleza o luz artificiales, y rarezas como dedos adicionales.
- **Busca marcas de agua** de generadores de imágenes comunes.
- Presta atención a los detalles de lo que se muestra: ¿es lógico? ¿Es apropiado?
- Al usar detectores, opta por una versión de alta resolución o la primera versión de una imagen en lugar de una que se ha compartido una y otra vez.