



MANUAL PRÁCTICO

DEEPFAKES Y MANIPULACIÓN DIGITAL

MTRO. EUGENIO JULIÁ TARNO

LAS FAKE NEWS

Las fake news o noticias falsas son contenidos que circulan principalmente en entornos digitales que presentan información falsa, distorsionada o engañosa como si fuera legítima y verificada.

A diferencia de un error humano o periodístico accidental, las fake news no surgen de una equivocación involuntaria. Generalmente son producidas y difundidas con una intención deliberada: influir en la opinión pública, afectar la reputación de una persona o institución, generar polarización social, obtener beneficios económicos mediante la monetización del tráfico digital o debilitar la confianza en actores públicos.

En contextos electorales, las fake news pueden utilizarse como herramienta para desacreditar candidaturas, cuestionar resultados, sembrar dudas sobre la integridad del proceso o erosionar la legitimidad institucional.

Las fake news suelen construirse sobre tres elementos fundamentales:

- **Apariencia de veracidad:** Utilizan formatos que imitan a los medios de comunicación serios (diseño web similar, logos parecidos, lenguaje periodístico) para ganar confianza rápidamente.
- **Intencionalidad:** Tienen el objetivo de perjudicar a alguien, beneficiar a un sector o simplemente crear caos y confusión.
- **Viralidad:** Se diseñan para apelar a las emociones (miedo, indignación o alegría extrema), lo que provoca que las personas las compartan de forma masiva en redes sociales antes de verificar si son ciertas.

Las fake news encuentran condiciones favorables en el ecosistema digital actual. Su rápida propagación responde a varios factores estructurales:

- **Apelan a emociones intensas**, lo que incrementa la interacción y el alcance algorítmico.
- **Simplifican problemas complejos**, ofreciendo explicaciones fáciles frente a fenómenos sociales o políticos difíciles de entender.
- **Refuerzan creencias previas**, aprovechando las cámaras de eco digitales donde los usuarios interactúan principalmente con contenidos alineados a sus posturas.
- **Se presentan como información urgente o exclusiva**, generando sensación de inmediatez que inhibe la verificación.
- **Circulan en entornos sin filtros editoriales inmediatos**, como grupos privados de mensajería o redes sociales descentralizadas.

TIPOLOGÍAS RECONOCIDAS

En el análisis contemporáneo del ecosistema digital, es fundamental distinguir entre tres categorías principales de fake news: información errónea, desinformación y malinformación. Aunque en la conversación pública suelen utilizarse como sinónimos, cada una responde a dinámicas distintas y requiere estrategias diferenciadas de atención y respuesta.

INFORMACIÓN ERRÓNEA



La información errónea se refiere a contenidos falsos o inexactos que son compartidos sin intención de causar daño. En este caso, quien difunde el mensaje cree genuinamente que la información es verdadera. Este fenómeno es frecuente en redes sociales y grupos de mensajería, donde usuarios replican noticias sin verificar su origen, especialmente cuando el contenido apela a la salud, seguridad o temas de interés colectivo.

DESINFORMACIÓN

La desinformación implica la creación y difusión intencional de información falsa con el objetivo de engañar, manipular percepciones o generar un beneficio político, económico o estratégico. A diferencia de la información errónea, aquí existe una voluntad explícita de distorsionar la realidad.



La fotografía de una madre descansando luego de varios días en el aeropuerto fue compartida en redes como señal de descuido.

MALINFORMACIÓN

La malinformación se refiere a información verídica que es utilizada de forma maliciosa o fuera de contexto para generar daño. No se trata de datos falsos, sino de contenidos auténticos manipulados en su narrativa, temporalidad o encuadre.

DEEPPFAKES

Los deepfakes son contenidos generados o manipulados mediante técnicas de inteligencia artificial, particularmente a través de modelos de aprendizaje profundo (deep learning), que permiten crear o alterar imagen, audio, video e incluso texto con un alto nivel de realismo.

Actualmente pueden manifestarse en cuatro formatos principales:

- **Texto:** Modelos generativos producen mensajes falsos con estilo y tono de figuras públicas, usados en campañas de desinformación.
- **Imagen:** Se crean rostros inexistentes o se manipulan fotos reales con gran realismo, dificultando la verificación.
- **Audio:** Síntesis de voz que imita timbre y entonación de personas reales, usado en fraudes telefónicos y suplantación.
- **Video:** Manipulación de gestos y expresiones faciales en clips, atribuyendo palabras o acciones falsas a funcionarios o celebridades.

[Audio aquí](#)

El 5G es un riesgo

La desinformación relacionada con el 5G abarca varios [mitos](#) que, por supuesto, son unos de los ejemplos de fake news más populares. Uno de ellos es la afirmación de que el 5G es perjudicial para la salud, a pesar de que numerosas investigaciones científicas, respaldadas por la OMS y fuentes como National Geographic, han confirmado que las ondas utilizadas por esta tecnología no representan un riesgo para los seres humanos.



¡Vamos campeón! Hoy es tu día para brillar, para demostrar de qué estás hecho. No te conformes con ser uno más, sé el mejor, el más fuerte. Recuerda que eres único, que tienes ese talento especial. ¡A por todas, siempre con actitud!



A diferencia de las fake news, que son contenidos informativos falsos creados para engañar o manipular, los deepfakes constituyen una herramienta tecnológica que puede utilizarse para producir contenido sintético. No toda fake news utiliza inteligencia artificial, pero un deepfake puede convertirse en soporte de una narrativa falsa.

Los deepfakes representan un desafío creciente debido a tres factores principales. Primero, el realismo creciente, que dificulta distinguir a simple vista entre contenido auténtico y contenido generado por IA. Segundo, la accesibilidad tecnológica, ya que existen aplicaciones y plataformas que permiten producir este tipo de material sin conocimientos técnicos avanzados. Y tercero, la velocidad de difusión, que posibilita que un contenido manipulado se viralice antes de ser verificado.

DINÁMICA DE PROPAGACIÓN DIGITAL

La desinformación no se viraliza de manera accidental; se expande dentro de un ecosistema digital donde interactúan tecnología, comportamiento humano e incentivos estratégicos. Comprender esta dinámica permite anticipar riesgos y definir respuestas proporcionales.

FACTORES ESTRUCTURALES DE PROPAGACIÓN

- **Algoritmos de recomendación:** Las plataformas priorizan contenidos con alta interacción. La información emocional o polémica suele generar más reacciones que los datos verificados, lo que facilita su amplificación.
- **Cámaras de eco:** La personalización del contenido expone a los usuarios principalmente a mensajes alineados con sus creencias. La repetición constante puede sustituir la verificación y generar percepción de consenso.
- **Reenvíos masivos:** En aplicaciones de mensajería privada, los contenidos pueden circular rápidamente sin filtros editoriales ni contexto adicional, dificultando su rastreo y verificación.
- **Baja alfabetización digital:** La falta de habilidades para evaluar fuentes o detectar manipulación incrementa la circulación involuntaria de información falsa.

ACTORES E INCENTIVOS

Las campañas de desinformación suelen involucrar múltiples actores: gobiernos, partidos políticos, grupos de presión, empresas privadas, influencers y cuentas automatizadas. Estos actores pueden actuar de manera coordinada o converger espontáneamente en narrativas similares.

Los incentivos que impulsan estas dinámicas incluyen:

- Beneficio político o electoral.
- Ganancias económicas derivadas del tráfico digital.
- Influencia en la agenda pública y percepción social.

AMPLIFICACIÓN Y HERRAMIENTAS DE MANIPULACIÓN

Además del comportamiento orgánico de los usuarios, existen mecanismos diseñados para acelerar la difusión:

- **Bots:** cuentas automatizadas que replican mensajes de forma masiva.
- **Granjas de contenido:** equipos organizados que producen y distribuyen narrativas coordinadas.
- **Microsegmentación:** uso de datos personales para dirigir mensajes específicos a grupos determinados, incrementando su efectividad.

INDICADORES TEMPRANOS DE RIESGO

La detección oportuna es fundamental para reducir el impacto de la desinformación. En entornos digitales, las narrativas manipuladas suelen mostrar patrones reconocibles antes de convertirse en crisis reputacional o tendencia dominante. Identificar estas señales permite activar monitoreo reforzado y evaluar la necesidad de intervención institucional.

SEÑALES DE ALERTA PRINCIPALES

- **Picos inusuales de menciones:** Un aumento repentino en el volumen de publicaciones sobre un tema, persona o institución, sin un evento que lo explique, puede indicar el inicio de una narrativa coordinada. Es importante revisar el origen de las menciones, el tipo de cuentas que amplifican y la plataforma donde surge la conversación.
- **Hashtags emergentes:** La aparición súbita de etiquetas asociadas a acusaciones o cuestionamientos puede reflejar intención de posicionar un encuadre específico. Debe analizarse quién lo impulsa y si existen mensajes repetitivos o migración rápida entre plataformas.
- **Cambios abruptos en sentimiento:** Variaciones significativas en el tono de la conversación, especialmente hacia una carga negativa dominante, pueden señalar activación coordinada.
- **Contenido audiovisual sin fuente verificable:** Videos o audios atribuidos a figuras públicas sin origen claro o respaldo institucional representan riesgo elevado y requieren verificación técnica inmediata.
- **Amplificación por cuentas nuevas:** La participación predominante de cuentas recientemente creadas, con mensajes idénticos o publicaciones simultáneas, puede indicar automatización o coordinación artificial.
- **Coordinación narrativa simultánea:** La repetición del mismo argumento o estructura discursiva en múltiples cuentas en corto tiempo es un indicador de posible campaña organizada.

EVALUACIÓN INTEGRADA DEL RIESGO

Para sistematizar esta evaluación y facilitar la toma de decisiones, se recomienda utilizar una matriz de clasificación de riesgo comunicacional

Indicador	Nivel Bajo	Nivel Medio	Nivel Alto
Volumen	Incremento leve y orgánico	Crecimiento sostenido	Pico abrupto sin evento que lo justifique
Actores	Cuentas orgánicas diversas	Influencers menores o cuentas amplificadoras	Cuentas nuevas, automatizadas o coordinadas
Contenido	Opiniones dispersas	Narrativa repetida	Mensajes idénticos o estructuralmente similares
Formato	Texto	Imagen con bajo alcance	Vídeo o audio sin fuente verificable

HERRAMIENTAS DE DETECCIÓN Y MONITOREO

La detección de desinformación y deepfakes no depende de una sola herramienta. Requiere la combinación de análisis técnico, monitoreo digital y criterio humano. La integración de estos elementos permite identificar riesgos con mayor precisión y reducir el margen de error en la toma de decisiones institucionales.

ANÁLISIS TÉCNICO AUDIOVISUAL

El análisis técnico se enfoca en examinar directamente el archivo digital para identificar señales de manipulación o generación sintética.

Las siguientes señales técnicas pueden indicar manipulación digital o generación sintética. Ninguna por sí sola confirma alteración, pero su combinación incrementa la probabilidad de contenido no auténtico.

Metadatos	Imagen	Audio	Vídeo
• Fecha de creación inconsistente • Ausencia de información del dispositivo • Modificaciones registradas en ediciones sucesivas • Incompatibilidad entre formato y origen declarado	• Sombras incoherentes con la fuente de luz • Bordes difusos o artefactos visuales • Reflejos imposibles o asimétricos • Pixelado desigual en zonas específicas	• Variaciones artificiales en tono o timbre • Ruido digital irregular • Ritmo de respiración inexistente o irregular • Cortes abruptos en frecuencia	• Desincronización entre labios y voz • Parpadeo poco natural o repetitivo • Movimientos faciales rígidos • Microexpresiones inconsistentes

HERRAMIENTAS DIGITALES ESPECIALIZADAS

Existen plataformas diseñadas específicamente para identificar contenido sintético o manipulado. Estas herramientas funcionan mediante modelos entrenados para detectar patrones característicos de generación por IA.

- **Detectores de contenido sintético:** Analizan patrones de píxeles, inconsistencias estadísticas o huellas digitales asociadas a modelos generativos.
- **Plataformas de verificación audiovisual:** Permiten examinar fotogramas, realizar búsquedas inversas de imagen y rastrear la circulación original del contenido.
- **Identificadores y marcas de agua digitales:** Algunas tecnologías incorporan marcas invisibles o firmas digitales que permiten rastrear si una imagen fue generada por inteligencia artificial. Sin embargo, no todos los modelos incluyen este tipo de identificación.

MONITOREO DE CONVERSACIÓN DIGITAL

Más allá del análisis del archivo, el monitoreo del entorno digital permite identificar patrones de amplificación y riesgo narrativo.

- **Social listening:** El monitoreo sistemático de menciones en redes sociales permite detectar picos anómalos, cambios de sentimiento y aparición de hashtags emergentes.
- **Identificación de redes coordinadas:** El análisis de interacción puede revelar cuentas que publican mensajes idénticos o similares en intervalos reducidos, lo que sugiere coordinación.
- **Detección de patrones automatizados:** Frecuencias de publicación inusuales, actividad 24/7 o replicación masiva son indicios de automatización.
- **Análisis de narrativas emergentes:** El seguimiento cualitativo permite identificar cómo evoluciona un encuadre discursivo, qué actores lo impulsan y si se desplaza hacia medios tradicionales.

INTEGRACIÓN DE INDICADORES: TÉCNICO, NARRATIVO Y DE RED

La detección efectiva de desinformación y deepfakes no debe basarse únicamente en el análisis del archivo digital ni exclusivamente en el monitoreo de redes sociales. Una evaluación integral combina tres dimensiones: técnica, narrativa y estructural (red).

El siguiente esquema permite visualizar cómo se articulan estos niveles de análisis:

Indicador Técnico	Indicador Narrativo	Indicador de Red
• Metadatos inconsistentes o ausentes • Inconsistencias visuales (sombras, bordes, parpadeo) • Desincronización labial o voz artificial • Patrones de compresión irregulares • Detección biométrica anómala	• Mensaje con acusaciones graves sin fuente verificable • Lenguaje emocional extremo o alarmista • Narrativa repetida con estructura idéntica • Uso de hashtags emergentes negativos • Enfoque dirigido a desacreditar institución o proceso	• Pico abrupto de menciones en corto tiempo • Amplificación por cuentas recién creadas • Publicaciones simultáneas o coordinadas • Migración rápida entre plataformas • Alta proporción de cuentas automatizadas

NOTA:

- Un solo indicador aislado no confirma manipulación.
- La presencia simultánea de señales técnicas, narrativas y de red incrementa la probabilidad de campaña coordinada.
- Cuando los tres niveles presentan anomalías, se recomienda activar evaluación de riesgo comunicacional y posible protocolo de respuesta.

EVALUACIÓN DEL RIESGO COMUNICACIONAL

La evaluación del riesgo comunicacional permite determinar cuándo una narrativa digital representa una conversación orgánica y cuándo constituye una amenaza potencial para la confianza institucional o la integridad del proceso electoral.

El objetivo no es reaccionar ante todo contenido crítico, sino identificar aquellos casos que, por su volumen, velocidad o encuadre, pueden escalar y afectar la percepción pública.

MATRIZ BÁSICA DE DECISIÓN

La siguiente matriz permite categorizar el nivel de riesgo y definir el tipo de acción institucional correspondiente:

Nivel de Riesgo	Características Observables	Acción Recomendada
Bajo	Bajo volumen de menciones, conversación dispersa, sin amplificación significativa.	Monitoreo continuo. No intervención pública inmediata.
Medio	Crecimiento sostenido, presencia en medios locales o influencers regionales, narrativa repetitiva.	Mensaje preventivo, aclaración técnica o posicionamiento breve.
Alto	Viralidad acelerada, cobertura en medios nacionales, afectación directa a la confianza institucional o acusaciones graves.	Respuesta oficial inmediata, vocería definida y activación de protocolo de crisis.

VARIABLES DE EVALUACIÓN

Para clasificar correctamente el nivel de riesgo, se recomienda analizar al menos cuatro dimensiones:

Variable	Preguntas clave	Indicador de Escalamiento
Volumen	¿Cuántas menciones se generan y en qué tiempo?	Incremento abrupto o tendencia exponencial.
Velocidad	¿En cuánto tiempo se posiciona la narrativa?	Escalamiento en pocas horas.
Actores	¿Quién amplifica el contenido?	Influencers, medios nacionales o cuentas coordinadas.
Contenido	¿Qué tipo de acusación se realiza?	Señalamientos de fraude, ilegalidad o manipulación institucional.

CRITERIOS PARA INTERVENIR

No toda crítica requiere respuesta institucional. La intervención debe activarse cuando se cumplan uno o más de los siguientes criterios:

- ¿Afecta la confianza en el proceso electoral?
- ¿Implica acusaciones de fraude o ilegalidad?
- ¿Proviene de actores con alta visibilidad o credibilidad?
- ¿Se está replicando simultáneamente en múltiples plataformas?
- ¿Existe evidencia de coordinación o automatización?

Cuando estos elementos coinciden, el riesgo trasciende la conversación digital y puede convertirse en crisis reputacional.

PRINCIPIO DE PROPORCIONALIDAD

La respuesta institucional debe ser proporcional al nivel de riesgo identificado:

- Riesgo bajo → Observación.
- Riesgo medio → Aclaración técnica o mensaje preventivo.
- Riesgo alto → Posicionamiento formal y vocería coordinada.

Sobrerreaccionar ante contenidos marginales puede amplificar innecesariamente la narrativa. Subestimar señales tempranas puede permitir su escalamiento.

PROTOCOLO DE RESPUESTA

La respuesta ante desinformación o deepfakes debe ser estratégica, coordinada y proporcional al nivel de riesgo identificado. Un protocolo claro reduce la improvisación, evita contradicciones públicas y protege la credibilidad.

El objetivo no es reaccionar ante todo contenido falso, sino intervenir de manera oportuna.

PRINCIPIOS CLAVE DE ACTUACIÓN

La respuesta debe regirse por cuatro principios fundamentales:

1. **Rapidez:** Responder antes de que la narrativa se consolide o migre a medios tradicionales. La velocidad es determinante en entornos digitales.
2. **Evidencia:** Sustentar cualquier posicionamiento con datos verificables, fuentes oficiales y respaldo técnico. La respuesta debe aportar claridad, no confrontación.
3. **Proporcionalidad:** No sobrerreaccionar ante contenidos marginales. Amplificar una narrativa menor puede otorgarle visibilidad innecesaria.
4. **Vocerías claras y coordinadas:** Definir una sola voz de respuesta. Mensajes fragmentados o contradictorios debilitan la credibilidad.

FASES DEL PROTOCOLO DE RESPUESTA

La actuación institucional puede organizarse en cuatro fases:



MATRIZ DE TIPO DE RESPUESTA

Nivel de Riesgo	Tipo de Respuesta	Canal Recomendado
Bajo	Monitoreo reforzado	Interno
Medio	Aclaración breve o ficha técnica	Redes oficiales
Alto	Comunicado formal y vocería pública	Conferencia, medios nacionales

QUÉ NO HACER

Una respuesta inadecuada puede amplificar la desinformación. Por ello se recomienda evitar:

- Repetir textualmente la fake news en encabezados o publicaciones.
- Responder de forma improvisada o emocional.
- Permitir vocerías múltiples no coordinadas.
- Ignorar narrativas que muestran señales claras de escalamiento.
- Descalificar sin evidencia técnica.

CONSIDERACIONES ESTRATÉGICAS ADICIONALES

- La primera respuesta debe enfocarse en datos, no en confrontación.
- El silencio puede ser estratégico en casos de bajo impacto.
- Es recomendable documentar cada caso para fortalecer protocolos futuros.
- La coordinación interinstitucional mejora la efectividad del mensaje.