Etiquetado del Contenido Generado por IA: Transparencia, Ética y Regulación

Sergi Bagó
Àlex Mitjans
Marc Sayós

Índice:

1. Introducción: El Desafío del Contenido Generado por IA

- La revolución del contenido digital:
  - Herramientas como ChatGPT, Midjourney, DALL·E y Synthesia permiten generar contenido realista en segundos, sin experiencia ni conocimientos necesarios.
- Riesgos emergentes:
  - Desinformación, manipulación, plagio, pérdida de autoría, “deepfakes”.
  - Impacto en política, educación, medios y redes sociales.
- El rol del etiquetado:
  - Informar al usuario sobre el origen sintético del contenido.
  - Promover transparencia, trazabilidad y confianza digital.
- Objetivo de este trabajo:
  - Evaluar el estado actual del etiquetado y su necesidad e implicaciones legales
  - Analizar tecnologías de etiquetado, su aplicación legal, social y sobretodo técnica.

2. ¿Qué es el etiquetado del contenido IA?

Añadir una indicación clara —visible o invisible— de que un contenido ha sido creado, total o parcialmente, mediante inteligencia artificial.

Objetivo principal

- Promover la transparencia
- Facilitar una interacción informada
- Proteger la confianza digital

Tipos de etiquetado según visibilidad

- **Visible**: Disclaimers como “Generado por IA”. Son fáciles de entender, pero pueden ser eliminados.
- **Invisible**: Metadatos, marcas de agua digitales. Son discretos y resistentes, pero manipulables si no se protegen bien.

Tipos según método

- **Manual**: El creador lo añade conscientemente.
- **Automático**: Lo incluye la herramienta por defecto (ej. ChatGPT, Firefly).
- **Mediante detección**: Algoritmos intentan identificar contenido IA a posteriori.

3. ¿Por qué es necesario etiquetar el contenido generado por IA?

- **Transparencia y confianza:** Permite a los usuarios tomar decisiones informadas y confiar en las plataformas.
- **Cumplimiento normativo:** Apoya leyes como el AI Act de la UE, garantizando trazabilidad y control del uso de IA.
- **Protección contra abusos:** Ayuda a frenar “deepfakes”, fraudes, manipulación y desinformación.
- **Ética en contextos sensibles:** Crucial en educación, periodismo y ciencia para preservar integridad, autoría y originalidad.

El etiquetado es más que una tecnología: es una herramienta de responsabilidad social en la era digital.

4. Tecnologías de Etiquetado e Identificación de Contenido IA

Las tecnologías se agrupan en **tres grandes categorías funcionales**:

1. **Watermarking en origen** → marcas invisibles incrustadas durante la generación.
2. **Atribución y trazabilidad** → metadatos estructurados, firmas digitales.
3. **Detección posterior** → análisis forense para identificar contenido IA ya distribuido.

En esta sección exploraremos:

- Visión general de cada tipo
- 3 tecnologías analizadas en profundidad:
    - **SynthID** (Google DeepMind)
    - **C2PA** (Adobe, Microsoft, BBC...)
    - **Polygraf AI** (detección textual con XAI)

4.1. Etiquetado en Origen (Watermarking)

¿Qué es?

Incrustar señales visibles o invisibles (marcas de agua digitales) directamente durante la **generación del contenido por IA**.

Características:

- Inserción en el espacio latente del modelo (imagen, texto, audio).
- Marcas persistentes, imperceptibles y redundantes.
- No afectan a la calidad visual ni textual.

Ventajas:

- Detectable incluso tras ediciones.
- Útil para trazabilidad inmediata tras la creación.

Limitaciones:

- Requiere acceso al modelo generativo.
- Puede ser eliminado si no se implementa correctamente.

Ejemplos:

- **SynthID (Google DeepMind)**
- steg.AI: marcas invisibles resistentes a fugas y edición.
- Adobe Firefly: metadatos y credenciales visuales incrustadas.

4.1. SynthID – Marca de Agua Invisible (Google DeepMind)

¿Qué hace?

- Incrusta señales invisibles en imágenes, vídeo y texto generados por IA.
- Permite su detección incluso tras ediciones (compresión, recortes, etc.).

Funcionamiento técnico

- Inserta marcas en el espacio latente del modelo generativo.
- Para texto, manipula probabilidades de tokens durante la generación.
- Señal redundante distribuida, resistente a transformaciones comunes.

Detección

- Modelo neuronal analiza el contenido y recupera la señal incrustada.
- No requiere el archivo original. Funciona con fragmentos.

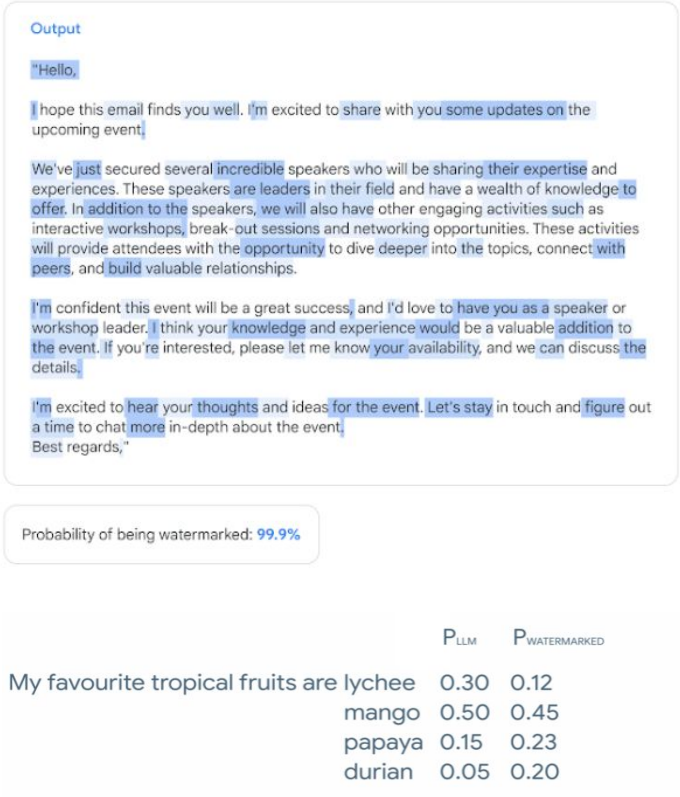
4.1. SynthID – Texto

¿Cómo se aplica la marca de agua?

- Durante la generación del texto, SynthID ajusta sutilmente las probabilidades de selección de palabras (tokens) en el modelo de lenguaje.
- Estos ajustes crean un patrón estadístico único que actúa como una firma digital, imperceptible para los lectores humanos pero detectable mediante herramientas especializadas .

¿Cómo se detecta la marca de agua?

- Se utiliza un detector bayesiano que analiza el texto en busca del patrón característico de la marca de agua.
- El detector puede clasificar el contenido en tres categorías: con marca de agua, sin marca de agua o incierto, dependiendo de la probabilidad de que el texto haya sido generado por IA .



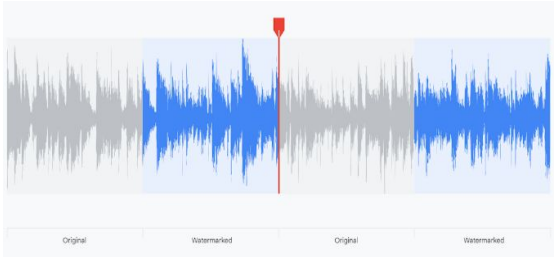
4.1. SynthID – Audio

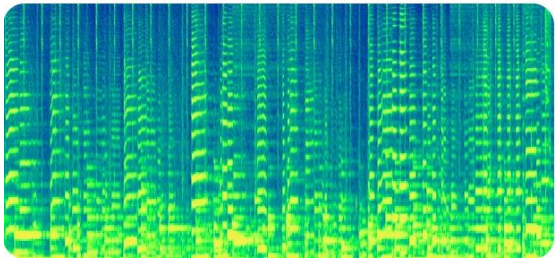
¿Cómo se aplica la marca de agua?

- Se convierte la onda de audio en un espectrograma (2d) STFT (Short-Time Fourier Transform)
- Se modifican ciertas energías del espectrograma. Estas modificaciones están diseñadas para:
  - Ser inaudibles para el ser humano (por debajo del umbral de percepción)
  - Ser estables frente a transformaciones comunes (como compresión MP3, ruido, pitch shift...)
  - Ser reconocibles mediante un detector que sabe lo que buscar
- Se convierte el espectrograma en la onda de audio (iSTFT).
- alectores humanos pero detectable mediante herramientas especializadas .

¿Cómo se detecta la marca de agua?

- Se convierte el audio sospechoso en un espectrograma.
- Se usa un modelo entrenado (como una CNN o una red bayesiana) que busca el patrón escondido en ese espectrograma.
- El modelo devuelve una probabilidad de que ese audio fue generado con SynthID.

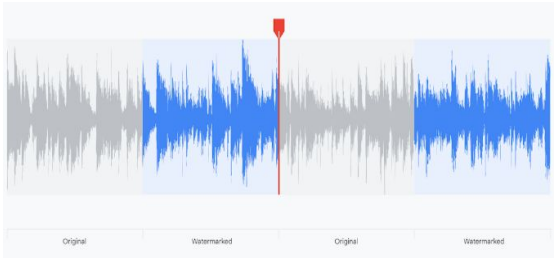




4.1. SynthID – Imagenes y Video

¿Cómo se aplica la marca de agua?

- Se aplican modificaciones sutiles a los píxeles de la imagen:
  - Se ajustan ligeramente los valores RGB de algunos píxeles (ejemplo: pasar de R=103 a R=104).
  - Estas modificaciones se distribuyen de forma espacial, como si se codificara un patrón binario (ej.: 00101011) en regiones específicas.
  - Se usan técnicas de parity encoding, frecuencia espacial y resiliencia al ruido para garantizar que el patrón siga presente incluso tras compresión, redimensionado, etc.



Si se modifican los píxeles de la imagen, ¿Por qué el ojo humano no lo detecta?

- El sistema visual humano es mucho más sensible al contraste y al color que a cambios pequeños en el valor absoluto de píxeles.
- SynthID aprovecha las limitaciones de la percepción humana para esconder la marca dentro de esos márgenes.
- Además, puede evitar modificar las regiones más sensibles (bordes definidos, caras, etc.) y centrarse en zonas de textura más uniforme donde los cambios se camuflan mejor.

4.2. Etiquetado para la Atribución: Metadatos, Firmas Digitales y Trazabilidad

Es la capacidad de **identificar quién creó un contenido**, **cómo ha sido modificado** y **si sigue siendo confiable**.

Criptografía aplicada a Metadatos, Firmas Digitales y Trazabilidad

Metadatos

- Uso de **manifiestos en JSON-LD** con metadatos.
- Los metadatos (EXIF, XMP) no están protegidos
- Solución: usar hashes criptográficos (ej: SHA-256)

Firmas Digitales (Autenticidad + No Repudio)

- Un autor firma digitalmente el hash del contenido + metadatos usando su clave privada.
- Algoritmos usados: ECDSA
- Verificación con clave pública: Llave pública se distribuye (o se obtiene de un certificado X.509 emitido por una autoridad de confianza)

Trazabilidad con cadenas de confianza

- Cada modificación puede generar una nueva firma que apunta a la anterior, formando una cadena de manifests.
- Esto permite seguir todo el historial con firmas encadenadas.

- **¿Quién creó este contenido?**
- **¿Cuándo se creó?**
- **¿Con qué herramientas?**
- **¿Fue modificado? ¿Por quién?**

Resultado:

Un archivo multimedia ya no solo es una imagen, sino un objeto verificable criptográficamente, con su historia y autoría protegidas matemáticamente.

Ejemplos:

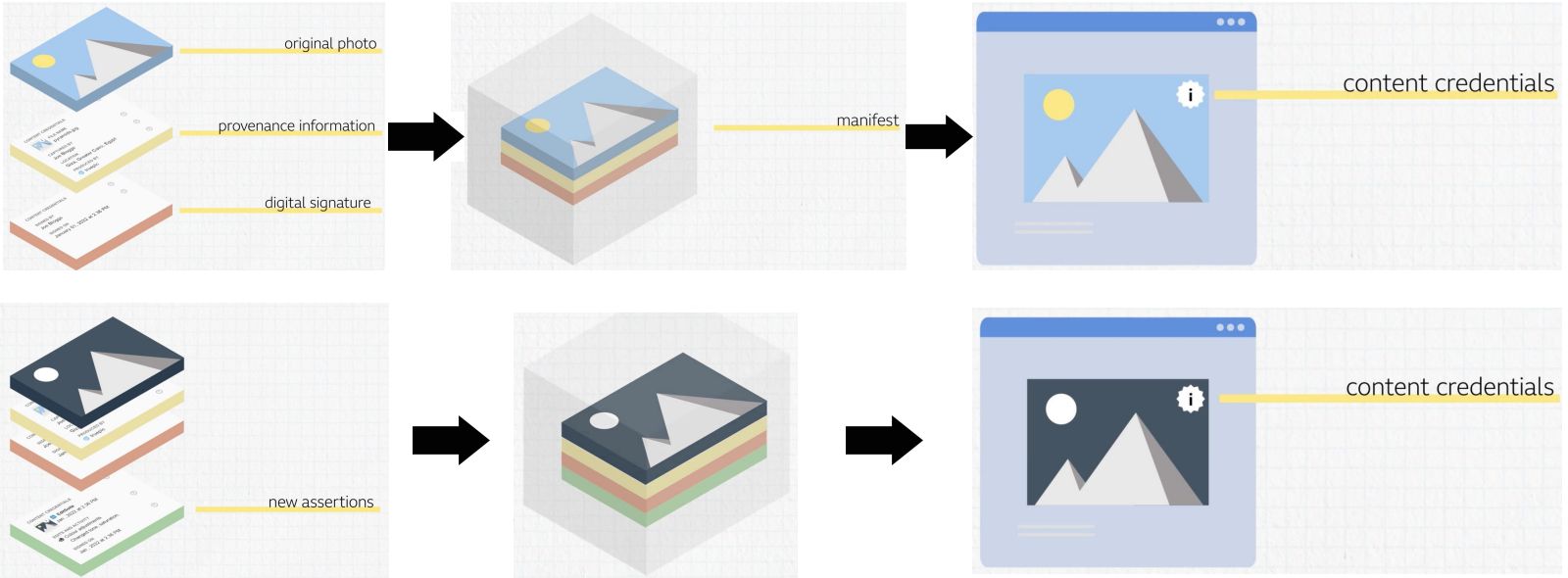
- **C2PA (Adobe, Microsoft, BBC)**
- Content Credentials: implementación práctica de C2PA en Adobe.
- W3C AI Content Labelling: metadatos semánticos web (JSON-LD, RDF).

4.2. C2PA – Coalition for Content Provenance and Authenticity

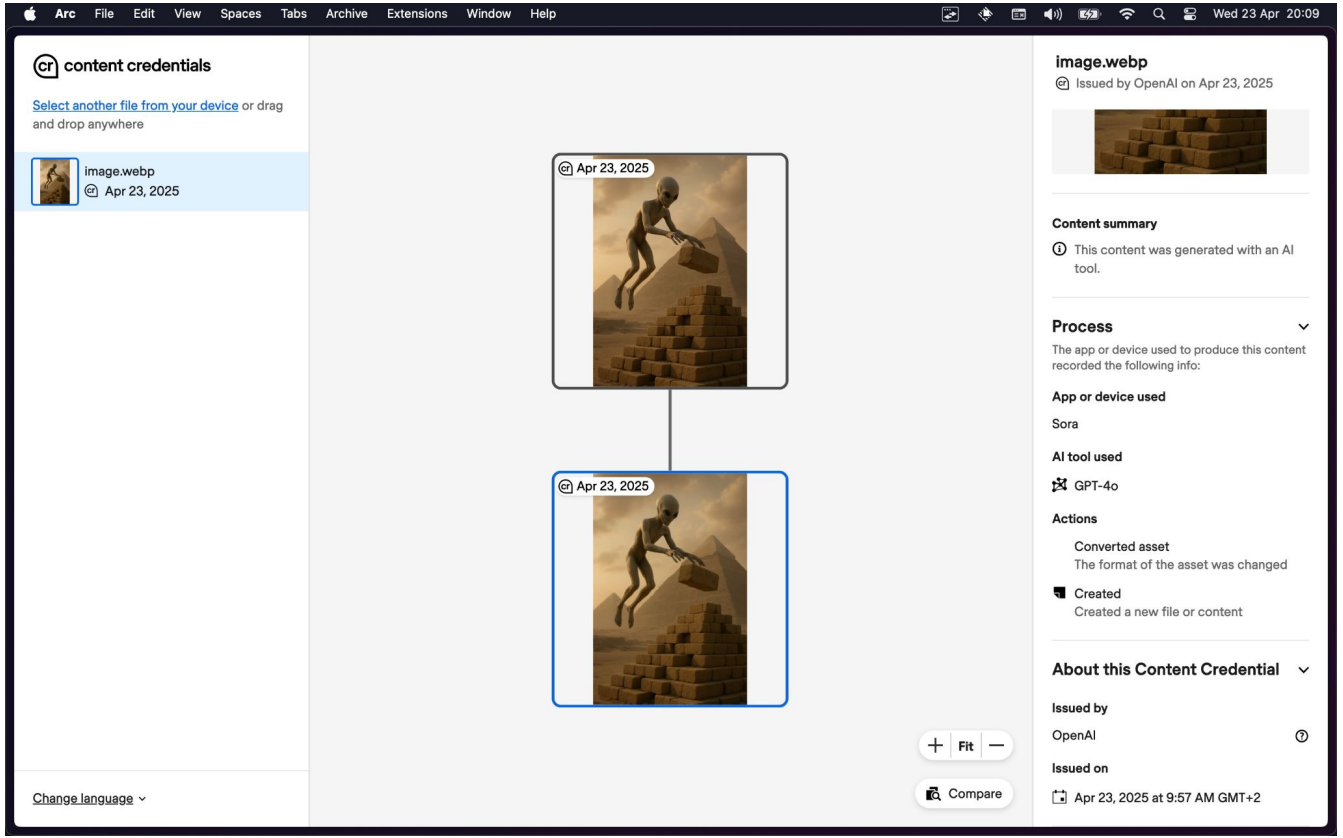
¿Qué es?

- Es un estándar abierto que permite asociar información de procedencia (provenance) a un archivo de medios (imagen, video, audio, documento). Esta información incluye quién creó el contenido, qué cambios ha sufrido, cuándo, y cómo. Todo queda registrado, firmado digitalmente y empaquetado de forma que cualquier persona o sistema pueda verificarlo.

¿Como funciona?



https://contentcredentials.org/verify -> podemos inspeccionar su contenido y ver como ha cambiado con el tiempo



4.3. Detección Posterior: Identificación de Contenido Generado por IA

¿Qué es?: Análisis técnico de contenido **ya publicado** para estimar si fue generado por IA,

Cómo funciona

- Modelos entrenados sobre grandes corpus de texto, imagen o audio.
- Detectan patrones sintéticos estadísticos, estilísticos o semánticos.
- Sistemas explicables (XAI) o clasificación directa.

Limitaciones

- Menor precisión frente a IA más sofisticada.
- Falsos positivos en contenido híbrido o reescrito.

Ventajas

- Aplicable a cualquier contenido, incluso sin marcas.
- Útil para regulación, auditoría y detección retroactiva.

Ejemplos: POLYGRAF GPTZero Copyleaks

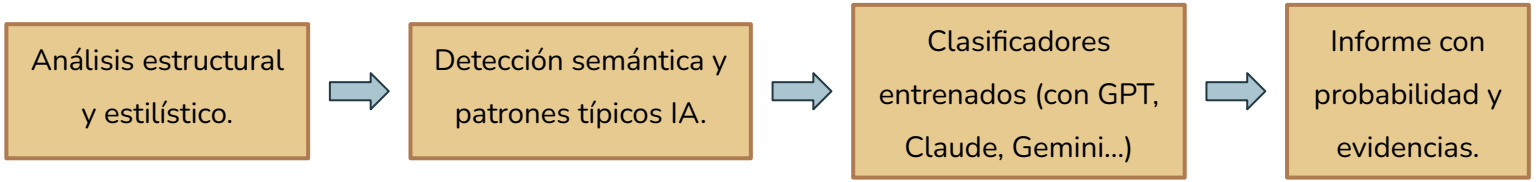
- Polygraf AI, GPTZero, Copyleaks, Originality.AI: detectores de texto IA.
- Vastav AI: análisis forense para deepfakes audiovisuales.
- Snorkel AI: generación de datos etiquetados con supervisión débil.

4.3. Polygraf AI – Detección Forense de Texto Generado por IA

¿Qué hace?

- Detecta texto generado por IA y explica por qué.
- Usa clasificación + XAI para detectar y justificar la decisión.

Flujo técnico



Ventajas

- Funciona con textos editados por humanos.
- API adaptable a entornos académicos, corporativos, etc.

5. Implicaciones Éticas del Etiquetado IA

Ética de la transparencia

- Ocultar el origen artificial del contenido se considera manipulación informativa.
- La transparencia protege la **autonomía del usuario** y promueve una interacción crítica y consciente.

Derechos de autor y responsabilidad

- En EE.UU., obras generadas únicamente por IA **no tienen protección legal** bajo copyright.
- Esto plantea dudas sobre **quién es el autor**, y **quién es responsable** ante usos indebidos o manipulaciones.

El etiquetado no es solo una medida técnica, sino un **compromiso moral con la veracidad**, la confianza pública y la rendición de cuentas en entornos digitales.

5.1. Marco Normativo Internacional

Unión Europea

- **AI Act (2024)**: obliga a etiquetar cualquier contenido sintético que pueda confundirse con real.

Estados Unidos

- **Ley federal propuesta (2023)**: divulgación obligatoria de contenido generado por IA.
- **California (AB 3211)**: marcas de agua obligatorias para deepfakes en elecciones.
- **Pensilvania (2024)**: notificación al consumidor al interactuar con IA.
- Otros estados (WA, FL, IL, NM): marcas de agua y herramientas de detección obligatorias.

China

- **Normas de 2025**: etiquetas visibles obligatorias + identificadores digitales (metadatos).
- Las plataformas son responsables de verificar y garantizar el cumplimiento legal.

Tendencia global: Todos los marcos coinciden en un objetivo común: **aumentar la transparencia, proteger al consumidor y mitigar el uso malicioso de contenido IA**.

6. Opinión Pública y Recepción

Preferencia por la transparencia

- El 65 % de los usuarios desea que se etiquete claramente el contenido generado por IA.

  (Pew Research Center, 2024)
- Mayor sensibilidad en contextos como noticias, política o campañas publicitarias.

Riesgos de banalización

- **Demasiadas etiquetas = pérdida de impacto.**
- "Fatiga de advertencias" → los usuarios ignoran mensajes repetitivos (como con cookies o contenido explícito).

Preocupaciones del público

- ¿Puede el etiquetado ser usado para **estigmatizar** o **limitar la creatividad**?
- ¿Genera una **falsa sensación de seguridad** asumir que lo no etiquetado es real?
- Mayor conciencia en regiones afectadas por desinformación.

Conclusión: El etiquetado debe ir acompañado de claridad, integridad y educación digital para ser efectivo.

7. Retos y Limitaciones

Diversidad tecnológica y falta de consenso

- No existe un estándar internacional unificado. Fragmentación legal y técnica.

Complejidad técnica

- Detección automática poco fiable frente a modelos IA cada vez más sofisticados

Manipulación y falsificación

- Riesgo de eliminar, modificar o falsificar etiquetas haciendo que el contenido sin etiquetar no garantice que sea auténtico.

Interoperabilidad limitada

- No todos los dispositivos/plataformas reconocen los mismos metadatos o marcas.

Dilemas éticos y creativos

- ¿Debe etiquetarse contenido parcialmente generado por IA? ¿Puede estigmatizar o limitar la creatividad?

8. El Futuro del Etiquetado IA

Hacia un ecosistema más confiable

- Métodos de identificación efectivos dependientes de los avances tecnológicos y con marcos legales sólidos

Detección Automática Inteligente

- Algoritmos que reconocen patrones sutiles de IA, incluso si se ocultan intencionadamente.
- Carrera tecnológica entre creadores de IA y sistemas de detección.

Estándar Global e Interoperabilidad

- Iniciativas como el **AI Act europeo** buscan unificar criterios de etiquetado.
- Beneficios de un estándar común:
  - Coherencia legal
  - Experiencia de usuario homogénea
  - Combate coordinado a la desinformación

Educación Digital y Conciencia Pública

- Informar al usuario sobre:
  - Qué significa una etiqueta
  - Cómo interpretarla
  - Por qué es relevante

9. Conclusión

IA: revolución y reto

- La IA ha transformado la creación digital: textos, imágenes, vídeos...
- Pero también plantea dudas sobre **autenticidad**, **autoría** y **responsabilidad**.

El rol del etiquetado

- Garantizar **transparencia y trazabilidad**.
- Tecnologías como **watermarking, metadatos y etiquetas visibles** ya están en uso.

Desafíos actuales

- Manipulación de etiquetas, falta de estándares, sobrecarga informativa.
- Necesidad de equilibrio entre **transparencia, privacidad y creatividad**.

El futuro del Etiquetado de IA

- El etiquetado no es un fin, sino parte de un **ecosistema de responsabilidad digital**.
- Solo con colaboración global podremos construir un entorno **ético, seguro y confiable**.

10. Preguntas

1. ¿Deberíamos exigir siempre que el contenido generado por IA esté etiquetado, incluso si apenas hay diferencia con el contenido humano?
2. Si un texto ha sido generado por IA y luego modificado por una persona, ¿aún debería llevar etiqueta? ¿Dónde trazamos el límite?
3. ¿Quién debe ser considerado autor de un contenido generado por IA: el modelo, el usuario o la plataforma?
4. ¿Deberían los contenidos generados por IA estar protegidos por derechos de autor? ¿Y si han sido solo asistidos por IA?
5. ¿Creéis que las etiquetas visibles realmente ayudan a los usuarios o se convertirán en "ruido" como los avisos de cookies?
6. ¿Hasta qué punto podemos confiar en herramientas de detección si incluso los expertos se equivocan con “deepfakes”?
7. ¿Creéis que el etiquetado puede estigmatizar injustamente a quien usa IA como herramienta creativa?
8. ¿Tiene sentido etiquetar contenido cuando el uso de IA está completamente normalizado y extendido?