

LA MARCA INVISIBLE DEL AI ACT

JON SANMILLÁN FUICA

Guía sobre cómo aplicar el artículo 50.2 al marcado y detección de contenidos generados o manipulados mediante IA

AI GOVERNANCE · EDICIÓN DE LA GUÍA: 10 DE SEPTIEMBRE DE 2026

RESUMEN. El artículo 50.2 del Reglamento de Inteligencia Artificial obliga a los proveedores de determinados sistemas de IA generativa a marcar sus resultados en un formato legible por máquina y a hacer posible su detección como contenidos generados o manipulados artificialmente. No exige simplemente insertar una etiqueta invisible; requiere construir una cadena técnica que permita incorporar una señal, conservarla, detectarla e interpretar el resultado con un nivel adecuado de efectividad, fiabilidad, robustez e interoperabilidad. Este artículo explica su ámbito, las excepciones, el modelo de cumplimiento propuesto por las Directrices de la Comisión y el Código de buenas prácticas, y analiza los casos documentados de Anthropic y Google para ilustrar tanto el potencial como los límites de estas soluciones.

CONTENIDO

1. Una obligación técnica, no una etiqueta visible.....	3
2. ¿Quién debe cumplir el artículo 50.2?	5
3. ¿Qué contenidos deben marcarse?	6
4. Las excepciones y los principales casos grises	8
5. El proceso de marcado ¿Cómo ofrecer una solución válida?	11
6. Cómo traduce el Código la obligación a la práctica.....	15
7. ¿Cómo abordan este tipo de mecanismos empresas como Anthropic y Google?	20
8. Un método de cumplimiento para proveedores	27
9. Aplicación temporal, supervisión y sanciones	32
10. Los límites del artículo 50.2	32
11. Conclusiones: una marca solo es útil si puede seguirse hasta una explicación	34
Referencias	35

Tres decisiones antes de actuar

1. Clasificar el sistema, las modalidades de salida y las excepciones antes de seleccionar una tecnología.
2. Diseñar una arquitectura de marcado y detección que siga siendo efectiva, fiable, robusta e interoperable durante la distribución.
3. Probar, documentar y actualizar la solución, comunicando sus límites sin convertir la ausencia de señal en prueba de autoría humana.

Límites. La guía no presenta una prueba empírica completa de las soluciones analizadas y la documentación pública de proveedores es parcial.

Copiamos un párrafo generado por un asistente de IA y lo publicamos en una red social. Descargamos una imagen creada con un generador y la reenviamos mediante una aplicación de mensajería. Recortamos unos segundos de un vídeo sintético, lo comprimimos y hacemos una captura de pantalla.

En los tres casos, la intervención de la inteligencia artificial puede resultar imperceptible para la persona que recibe el contenido. También puede desaparecer parte de la información técnica asociada al archivo durante su recorrido. Un metadato puede eliminarse al subir una imagen a una plataforma; una marca de agua puede degradarse al comprimir un vídeo; y un patrón estadístico insertado en un texto puede perderse después de una reescritura profunda.

Esta es la dificultad que pretende abordar el [artículo 50.2 del Reglamento de Inteligencia Artificial](#). El precepto no se limita a ordenar que el contenido lleve una marca. Exige que los resultados de los sistemas de IA afectados se encuentren **marcados en un formato legible por máquina y que puedan ser detectados como generados o manipulados artificialmente**.

Las dos partes son inseparables. Una marca que nadie puede comprobar ofrece una transparencia incompleta. Un detector que no sabe qué señal debe buscar tampoco resuelve el problema. La obligación solo adquiere sentido cuando existe una cadena funcional:

marcar el contenido → conservar la señal → ofrecer un medio de detección → presentar un resultado comprensible.

Las [Directrices de la Comisión sobre las obligaciones de transparencia del artículo 50](#), publicadas el 20 de julio de 2026, y el [Código de buenas prácticas sobre el marcado y etiquetado de contenidos generados mediante IA](#) convierten esa formulación general en un modelo mucho más concreto. Aclaran qué resultados quedan incluidos, qué transformaciones pueden quedar fuera, cómo debe facilitarse la detección y qué soluciones considera actualmente adecuadas la Comisión.

La tesis de este artículo

Mi lectura es que **el artículo 50.2 no establece solamente una obligación de insertar marcas de agua**. Establece una obligación de construir una infraestructura de transparencia. El proveedor debe poder justificar que:

1. Ha identificado el tipo de contenido que genera la IA que queda sujeto al precepto.
2. Ha incorporado una señal técnica adecuada a cada modalidad.
3. Ha previsto cómo conservarla a lo largo de la cadena de valor.
4. Ha facilitado un medio para localizarla.
5. Ha explicado el resultado sin atribuir a la detección más valor del que realmente tiene.
6. Ha probado y mantenido todo el sistema durante su ciclo de vida.

Esta tesis guía el resto del análisis. Las Directrices aportan el perímetro jurídico; el Código propone una arquitectura operativa; y los casos de Anthropic y Google permiten comprobar hasta dónde llega actualmente la evidencia pública sobre su implantación.

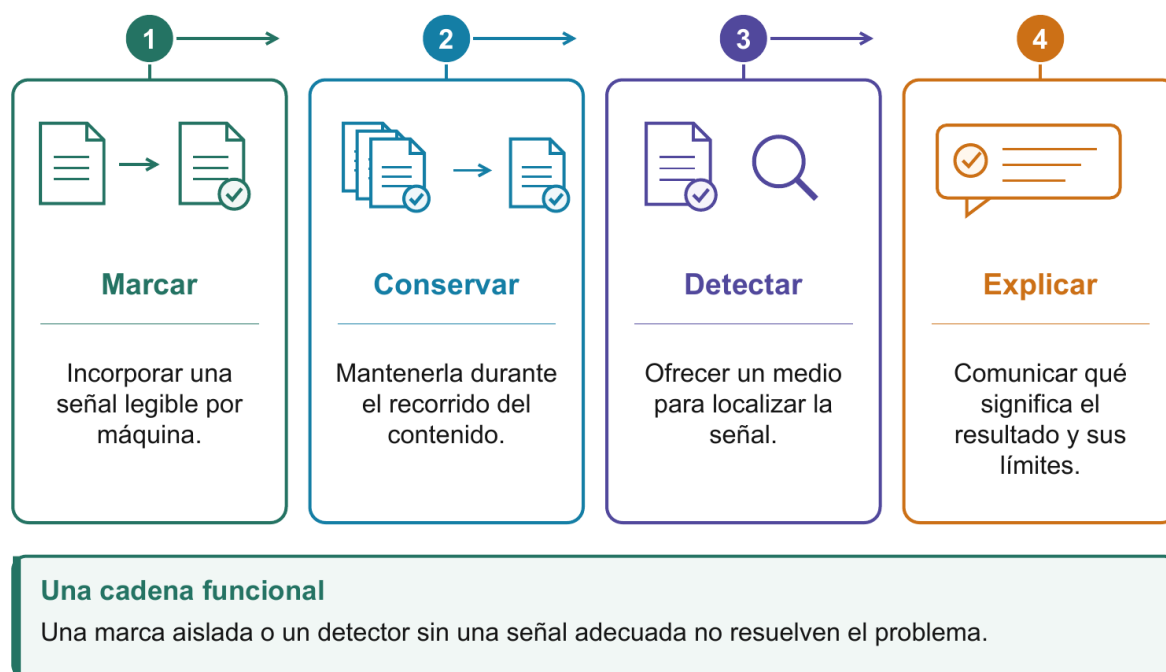


FIGURA 1. La cadena técnica de transparencia. La señal debe poder seguirse desde su inserción hasta una explicación comprensible. Síntesis de la introducción.

1. UNA OBLIGACIÓN TÉCNICA, NO UNA ETIQUETA VISIBLE

El artículo 50.2 se dirige a los **proveedores de sistemas de IA, incluidos los sistemas de IA de uso general, que generen contenidos sintéticos de audio, imagen, vídeo o texto**. Estos proveedores deben asegurar dos resultados:

1. Que las salidas se encuentren marcadas en un formato legible por máquina.
2. Que sea posible detectarlas como generadas o manipuladas artificialmente.

Además, las **soluciones técnicas deben ser efectivas, interoperables, robustas y fiables** en la medida en que sea técnicamente viable, atendiendo a las particularidades del

contenido, las limitaciones de cada modalidad, los costes de implantación y el estado de la técnica generalmente reconocido.

A este respecto, es importante realizar la primera distinción: **una marca legible por máquina no tiene por qué ser perceptible para una persona**. Puede consistir en metadatos firmados, una marca de agua imperceptible o un patrón estadístico que solo una herramienta técnica sea capaz de reconocer.

Por ello, el artículo 50.2 no debe confundirse con el artículo 50.4, que analicé en el artículo anterior. Ambos apartados forman parte de la misma arquitectura de transparencia, pero responden a preguntas distintas:

Tabla 1. Comparación entre el marcado técnico del artículo 50.2 y la divulgación perceptible del artículo 50.4

Elemento	Artículo 50.2	Artículo 50.4
Sujeto obligado	Proveedor del sistema de IA generativa	Responsable del despliegue que utiliza el sistema
Objeto	Audio, imagen, vídeo o texto sintético cubierto por el precepto	Ultrafalsificaciones y determinados textos de interés público
Forma principal	Marcado legible por máquina y medio de detección	Divulgación perceptible para la audiencia
Finalidad inmediata	Permitir verificar técnicamente el origen artificial	Informar directamente a la persona expuesta
Visibilidad	Puede ser imperceptible	Debe ser clara y distinguible

Un metadato invisible no sustituye la divulgación que debe realizar el responsable del despliegue cuando publica una ultrafalsificación. Del mismo modo, escribir "contenido creado con IA" debajo de una imagen no exonera al proveedor de incorporar el marcado técnico y facilitar su detección cuando resulte aplicable el artículo 50.2. Las obligaciones pueden, por tanto, acumularse sobre un mismo contenido y corresponder a sujetos distintos. El proveedor prepara la infraestructura técnica; el responsable del despliegue informa a la audiencia en los casos previstos por el artículo 50.4; y la persona expuesta puede acudir a un mecanismo de detección para comprobar el origen del resultado.

1.1. ¿Qué problema pretende resolver?

El considerando 133 parte de una premisa fácil de observar: cada vez resulta más difícil distinguir el contenido sintético del contenido creado o registrado sin IA. Esa indistinguibilidad puede facilitar **la desinformación, la manipulación, el fraude, la suplantación y el engaño a los consumidores**.

Para reducir esos riesgos, el Reglamento menciona diferentes técnicas que pueden ayudar a identificar la procedencia artificial: marcas de agua, metadatos, métodos criptográficos de autenticidad y procedencia, registros o huellas digitales. La elección no se agota en una tecnología concreta. Lo relevante es que la solución funcione como un sistema y satisfaga los requisitos de calidad del precepto.

El considerando 135 añade una segunda idea: esta transparencia depende de la cooperación a lo largo de la cadena de valor. De poco sirve que el proveedor incorpore metadatos si las plataformas los eliminan automáticamente al procesar el archivo. Tampoco basta con insertar una marca propietaria si ninguna persona o servicio externo puede localizar el detector correspondiente.

El considerando 136 conecta estas obligaciones con el ecosistema informativo y con las medidas del Reglamento de Servicios Digitales. **El marcado puede ayudar a plataformas, investigadores, medios, verificadores y autoridades a identificar contenido sintético y a adoptar medidas frente a riesgos sistémicos** en ámbitos como el de las elecciones de un país. Por ejemplo, *Meta utiliza indicadores técnicos compartidos por el sector, como metadatos C2PA e IPTC, además de la declaración del usuario*, para aplicar etiquetas sobre determinados contenidos publicados en Facebook, Instagram y Threads. Esta práctica no equivale a detectar cualquier marca de agua ni todo contenido generado mediante IA.

Dicho esto, es imprescindible tener en cuenta que **identificar el origen artificial de un contenido no determina su licitud ni su veracidad**. Una imagen marcada puede ser completamente verídica en su mensaje. Un vídeo no marcado puede ser una grabación real y, al mismo tiempo, estar presentado de forma engañosa. Un texto generado por IA puede contener información correcta, mientras que un texto escrito por una persona puede ser falso. El artículo 50.2 no crea una máquina de la verdad. Crea una infraestructura para aportar información sobre **cómo se originó o transformó el contenido**.

Es importante tener en cuenta de que esta obligación persigue afrontar este problema, permitiendo a los proveedores establecer **un marcado que no empeore la experiencia de usuario como lo puede hacer el etiquetado visible del contenido**. A menudo las marcas de agua visibles pueden generar rechazo entre los consumidores, el contenido marcado pierde calidad. Este precepto pretende solventar el problema de la desinformación sin la necesidad de que se produzcan estos efectos negativos sobre el contenido generado

2. ¿QUIÉN DEBE CUMPLIR EL ARTÍCULO 50.2?

2.1. El proveedor del sistema

El sujeto obligado es el **proveedor del sistema que genera o manipula el contenido**. Conforme al artículo 3 del AI Act, es proveedor **quien desarrolla un sistema de IA, o encarga su desarrollo, y lo introduce en el mercado o lo pone en servicio bajo su propio nombre o marca**, de forma remunerada o gratuita.

La obligación no se limita a las empresas que entrenan grandes modelos fundacionales. También puede alcanzar a un operador que integra un modelo de un tercero en un producto propio y comercializa el sistema resultante bajo su marca. La calificación dependerá de la posición real del actor en la cadena de valor y no de la denominación que utilice en el contrato.

Las Directrices permiten que un **proveedor se apoye en marcas y detectores implementados por un proveedor situado en una fase anterior o por un tercero especializado**. Sin embargo, esa externalización técnica **no desplaza su responsabilidad**. Quien coloca el sistema bajo su nombre debe comprobar que la solución heredada cubre las modalidades, transformaciones y condiciones de uso de su propio producto.

Caso práctico. Una empresa construye un generador de imágenes sobre un modelo de un tercero. El modelo base incorpora una marca de agua, pero la aplicación aplica después un filtro y exporta la imagen en un formato diferente. La empresa no debería limitarse a mencionar que el modelo de origen ya estaba marcado: debe verificar que su proceso no elimina la señal y que existe un mecanismo capaz de detectar el resultado final. Microsoft 365 Copilot, por ejemplo, integra modelos fundacionales de Anthropic y

OpenAI y, a pesar de no ser suyos los modelos, debe asegurarse de que los outputs (dentro del ámbito del art. 50.2) generados por Copilot se encuentran marcados aunque se generen con los modelos de Anthropic u OpenAI.

2.2. Sistemas de propósito específico, de uso general y agentes

Las Directrices interpretan el ámbito de forma funcional. Puede incluir:

- Sistemas diseñados específicamente para generar una modalidad, como una herramienta de síntesis de voz.
- Sistemas de IA de uso general capaces de producir textos, imágenes, audio o vídeo.
- Sistemas multimodales que combinan varias de estas salidas.
- Sistemas agénticos cuando generan contenido sintético destinado a ser percibido por personas.

No todo resultado interno de un agente queda incluido. Una solicitud técnica entre servicios, una llamada a una API o un paso intermedio que no está destinado a ser percibido por personas no se convierte automáticamente en "contenido" sujeto a marcado. En cambio, el correo, informe, imagen o respuesta final que el agente produce para una persona puede quedar comprendido.

2.3. Proveedores establecidos fuera de la Unión

El ámbito territorial del AI Act puede alcanzar a proveedores establecidos fuera de la Unión **cuando introducen el sistema en el mercado europeo, lo ponen en servicio en la Unión o se cumplen los restantes criterios territoriales del Reglamento**. Una arquitectura global no permite asumir que el artículo 50.2 puede aplicarse exclusivamente a los servidores europeos si el mismo sistema genera contenidos que previsiblemente circularán en la Unión. Ahora bien, **la utilización meramente incidental, imprevisible o no autorizada de un resultado en la Unión no basta por sí sola para someter a un proveedor de un tercer país que no haya dirigido el sistema al mercado europeo ni lo haya puesto en servicio en la Unión**; el vínculo territorial debe analizarse conforme al artículo 2 y a las Directrices.

3. ¿QUÉ CONTENIDOS DEBEN MARCARSE?

Las Directrices proponen comprobar de forma acumulativa cuatro elementos:

1. Existe un **sistema de IA** en el sentido del Reglamento.
2. El sistema **genera o manipula contenido sintético**.
3. El resultado consiste en **audio, imagen, vídeo o texto**, o en una combinación de estas modalidades.
4. No concurre una de las excepciones previstas.

La obligación puede alcanzar **tanto a contenido completamente generado por IA como a contenido humano parcialmente manipulado**. No se exige que la aportación artificial sea exclusiva. Tampoco se limita a los contenidos que constituyen una ultrafalsificación ni a los textos de interés público como el art. 50.4.

3.1. Audio, imagen, vídeo y texto

Las cuatro modalidades se interpretan atendiendo a su capacidad para transmitir contenido perceptible por personas:

- **Audio:** voz, música, efectos, paisajes sonoros o combinaciones de ellos.
- **Imagen:** fotografías, ilustraciones, gráficos o representaciones visuales sintéticas.
- **Vídeo:** secuencias audiovisuales, animaciones y contenido en movimiento.
- **Texto:** prosa, artículos, mensajes, informes u otros contenidos lingüísticos destinados a personas.

Las Directrices incluyen **también resultados multimodales y contenidos perceptibles en entornos de realidad virtual, aumentada o mixta**. Un objeto tridimensional o un gemelo digital puede quedar comprendido cuando su presentación se traduce en una experiencia visual o audiovisual destinada a una persona.

3.2. Contenido mixto

El artículo 50.2 puede aplicarse aunque el **resultado combine aportaciones humanas y artificiales**. Lo interesante no es si una IA ha realizado el cien por cien de la obra, sino si el sistema ha generado o manipulado contenido sintético cubierto por el precepto. Añadir una frase humana a un texto generado por IA o retocar manualmente una imagen no elimina por sí solo la obligación. Tampoco lo excluye que el sistema IA edite un contenido que tú le aportas (a no ser, como ya se verá, que la edición sea muy limitada). Habrá que valorar la salida del sistema y la transformación posterior, especialmente si la marca debe sobrevivir al flujo de edición.

3.3. Resultados que las Directrices dejan fuera

La expresión "texto, imagen, audio o vídeo" podría interpretarse de forma ilimitada. Las Directrices evitan esa consecuencia y **excluyen diversos resultados** que no cumplen la finalidad del precepto:

- La **mera reproducción, presentación, clasificación o recomendación de contenido existente**, cuando el sistema no lo genera ni lo manipula.
- **Datos procedentes de sensores, telemetría o transmisiones técnicas** que no están destinadas a ser percibidas como contenido por personas.
- **Cadenas breves de números, símbolos o letras, palabras aisladas, etiquetas de interfaz, pies automáticos muy breves o gráficos reducidos a la escala de un icono**.
- **Código fuente, lenguajes de configuración, consultas SQL, archivos YAML o JSON, SDK y llamadas a API dirigidas a sistemas técnicos**.
- **Comunicaciones de máquina a máquina y pasos intermedios de un proceso cerrado que no van a exponerse a una persona**.

Cuando un proceso industrial genera cientos de imágenes intermedias para producir una pieza final, las Directrices permiten **concentrar el marcado en el resultado que vaya a salir del entorno cerrado** y llegar a personas. La obligación no pretende marcar cada operación interna sin relevancia comunicativa.

Tabla 2. Supuestos orientativos sobre resultados incluidos y excluidos del artículo 50.2

Supuesto	Resultado orientativo
Un asistente redacta una nota de prensa completa	Incluido, salvo que concurra una excepción
Un generador crea una imagen publicitaria fantástica	Incluido aunque no sea una ultrafalsificación
Una IA ordena artículos ya publicados sin modificarlos	Fuera: no genera ni manipula el contenido

Supuesto	Resultado orientativo
Un agente envía una consulta JSON a otra herramienta	Fuera: comunicación técnica entre sistemas
Una aplicación genera un informe final a partir de sensores	El informe puede quedar incluido; los datos brutos, no necesariamente
Una herramienta propone una palabra aislada para un botón	En principio, fuera conforme a las Directrices

3.4. Regímenes proporcionales para entornos cerrados y contenidos efímeros

Además de los casos que quedan expresamente fuera del artículo 50.2, las Directrices identifican determinados supuestos estrechos en los que los costes del marcado pueden resultar desproporcionados frente a su utilidad para la transparencia.

- En primer lugar, cuando un **sistema generativo se encuentra integrado en un producto físico** que opera en un entorno técnicamente controlado y principalmente instructivo, puede considerarse suficiente una solución basada en metadatos menos robustos. Para ello deben existir medidas eficaces que impidan que el resultado sea capturado, exportado o difundido fuera de ese entorno.
- En segundo lugar, **determinados usos industriales o entre empresas** pueden no requerir marcado ni detección cuando concurren acumulativamente tres condiciones: que el resultado sea **estrictamente técnico**; que solo vaya a ser **percibido por un número limitado y predeterminado de profesionales de las organizaciones implicadas**; y que **no esté destinado a salir de ese entorno**, existiendo salvaguardas eficaces frente a una difusión previsible.
- Finalmente, el **contenido generado en tiempo real, consumido de forma inmediata y no grabado, almacenado ni difundido posteriormente** (por ejemplo, ciertas experiencias de videojuegos o realidad virtual) puede quedar exceptuado cuando el marcado no sea técnicamente viable y las personas expuestas sean informadas mediante un aviso dentro de la experiencia o de la sesión.

Estas precisiones proceden de la interpretación no vinculante de la Comisión sobre **proporcionalidad**. No deben convertirse en una excepción general para cualquier uso interno, profesional o B2B.

4. LAS EXCEPCIONES Y LOS PRINCIPALES CASOS GRISES

El propio artículo 50.2 nos dice que "esta obligación no se aplicará en la medida en que los **sistemas de IA desempeñen una función de apoyo a la edición estándar o no alteren sustancialmente los datos de entrada** facilitados por el responsable del despliegue o su semántica, o cuando estén autorizados por ley para detectar, prevenir, investigar o enjuiciar delitos". Estas excepciones deben interpretarse a partir del resultado real, no de la etiqueta comercial de la herramienta.

4.1. Edición estándar

No se exige el marcado cuando el sistema se limita a realizar una función de asistencia para una **edición estándar**. Las Directrices determinan que se produce una edición estándar cuando el objetivo es preparar el contenido existente, no generar nuevo

contenido. Si se produce un cambio material en el contenido que afecte a su significado, estilo o intención, la edición quedará sujeta a la obligación de marcado. Se sitúan aquí operaciones como:

- **Ajustes menores de legibilidad o formato.**
- **Compresión de archivos o cambios técnicos de formato.**
- **Reducción de ruido, estabilización limitada o mejora técnica de una grabación.**
- **Recortes, ajustes de color, luz o nitidez que no alteran el significado de una imagen.**
- **Eliminación de polvo, ojos rojos o desenfoque de un fondo por privacidad.**

4.2. Ausencia de alteración sustancial

También quedan fuera los sistemas que **no alteran sustancialmente los datos de entrada aportados por el responsable del despliegue ni su semántica**. Esta valoración es contextual.

Las Directrices sitúan, en principio, dentro de esta zona de exclusión las **transcripciones, la corrección ortográfica y gramatical, determinadas funciones de accesibilidad, las traducciones que preservan el significado y otros cambios técnicos que no modifican el mensaje**. En cambio, suelen exigir marcado:

- Los resúmenes que seleccionan y condensan información.
- Las paráfrasis o reescrituras que cambian el estilo, la estructura o el sentido.
- La sustitución o incorporación de personas y objetos con efectos semánticos.
- La síntesis de voz o la fabricación de acontecimientos audiovisuales.
- Las composiciones que combinan elementos para construir una representación nueva.

La frontera no siempre será evidente. Una traducción jurídica literal puede conservar el significado, mientras que una localización creativa puede alterar el tono, las referencias y la intención del texto. Por su parte, un corrector puede limitarse a reparar errores o reescribir por completo la voz del autor. El proveedor deberá analizar las capacidades y usos previstos de su sistema, y no solo el nombre de la función. Nuevamente, la clave para determinar cuándo se produce una alteración sustancial de los datos de entrada, será imprescindible tener en cuenta aspectos como **el formato; el tipo de contenido; o si afecta al significado, estilo o intención del contenido**.

4.3. Uso autorizado para fines de aplicación de la ley

La obligación tampoco se aplica cuando el sistema se encuentra legalmente autorizado para detectar, prevenir, investigar o perseguir infracciones penales. No se trata de una excepción general para cualquier organismo público o actividad de seguridad. Deben concurrir la autorización legal y las finalidades concretas previstas por el Reglamento.

4.4. Cumplir más allá del mínimo

Que un contenido quede fuera de la obligación no impide marcarlo voluntariamente. Anthropic, por ejemplo, ha anunciado que las traducciones generadas por los futuros modelos de Claude también incorporarán su marca de agua, aunque las Directrices contemplen traducciones no sustanciales dentro de los posibles supuestos excluidos.

Esta diferencia no implica necesariamente una contradicción. Un proveedor puede adoptar una regla técnica más amplia y uniforme si la explica correctamente. De hecho, para los proveedores de estos sistemas IA es mucho más cómodo introducir el marcado en la totalidad de este tipo de contenidos (aunque queden fuera del ámbito de la obligación),

que tener que establecer procedimientos para determinar qué tipo de contenido queda marcado y cuál no. Dicho esto, surge un problema primordial y es que **el marcado voluntario puede transmitir una conclusión equivocada sobre la atribución social de la obra o sobre el grado de intervención de la IA en determinados contenidos**. Los marcados genéricos merman la percepción social del mérito humano de los verdaderos autores. Para ser sinceros, el hecho de que ya esté marcado como contenido generado o manipulado con IA, ya lo reduce, aunque se especifique que se hayan realizado únicamente retoques menores. Esta es una de las principales tensiones que genera este artículo.

Mi criterio: la solución más práctica y sostenible es el marcado general de todo el contenido generado, aunque tenga consecuencias negativas en la atribución social del contenido o la percepción del mérito humano por parte de la sociedad. Las Directrices incluyen entre los posibles supuestos no sustanciales una traducción que preserve el significado. Anthropic, en cambio, ha anunciado que también marcará las traducciones producidas por los futuros modelos de Claude porque todas las palabras han sido seleccionadas por el sistema. La dificultad (técnica y económica) para determinados tipos de contenido de establecer procedimientos internos en estos sistemas con el objetivo de determinar qué contenido necesita estar marcado y qué contenido queda fuera del ámbito de esta obligación, causa que las empresas se orienten al etiquetado de todo tipo de contenido sintético generado por ellas. Además de marcar todo el contenido generado, me temo que, en muchos casos, el marcado se llevará a cabo de forma genérica, sin explicar correctamente qué significa su señal y hasta dónde alcanza en cada caso. Dicho esto, personalmente, creo que es la solución que resulta más práctica y sostenible para los proveedores. Marcados personalizados de cada resultado o procedimientos internos del sistema que determinen si procede el marcado o no, producirían un aumento considerable en el tiempo de generación del contenido, en su precio y demandan actuaciones de integración técnica más complejas. El coste de esta transparencia y de la información mediante el marcado es caer en el riesgo de que el receptor equipare "intervención de la IA" con "autoría de la IA", más aún cuando se realice sobre contenidos en los que apenas intervino. Por ello empresas como Anthropic han sufrido grandes críticas tras la publicación de sus procesos y técnicas de marcado.

Para mí este asunto está ligado con la percepción negativa que la sociedad tiene de todo el contenido generado por IA. Generalmente se desvaloriza el trabajo del autor que utiliza este tipo de herramientas. No digo que esto no tenga sentido en algunos casos, hay mucho contenido basura, pero creo que la sociedad debe de ser capaz de tener una opinión crítica más profunda a la hora de realizar estos juicios y no limitarse al rechazo sistemático de este tipo de contenido

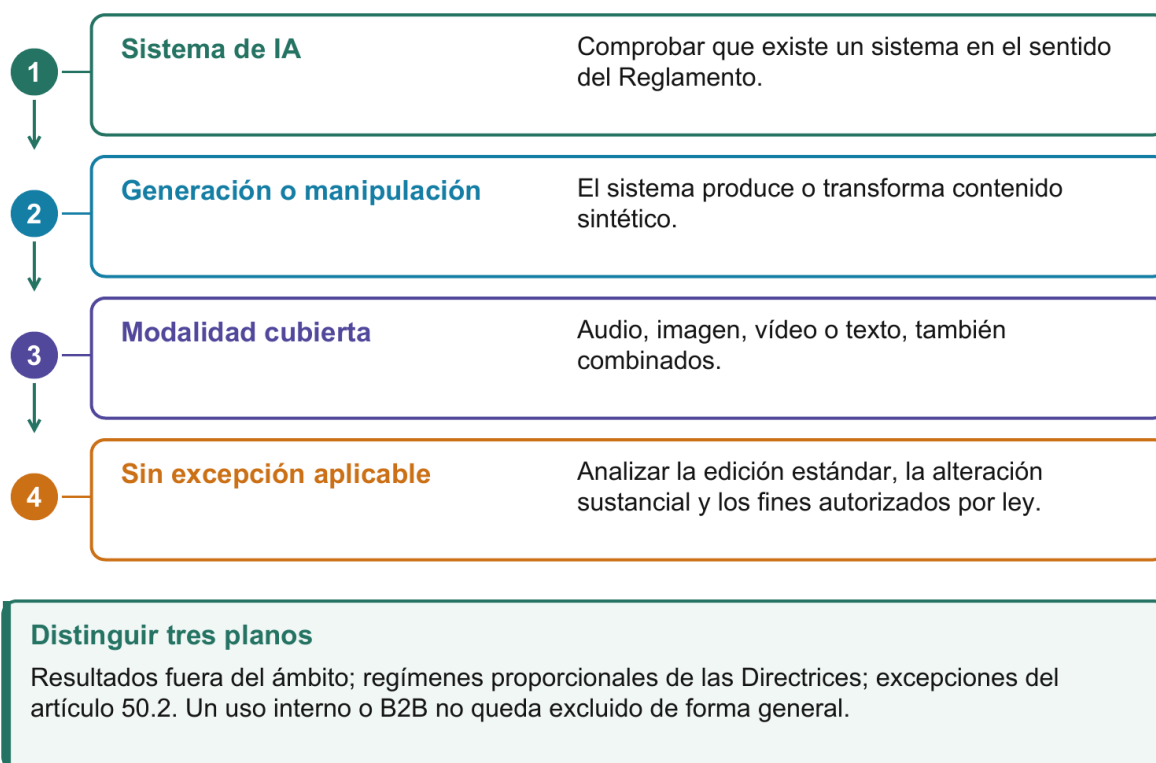


FIGURA 2. Esquema orientativo para determinar el ámbito del artículo 50.2 antes de seleccionar la solución de marcado y detección.

5. EL PROCESO DE MARCADO ¿CÓMO OFRECER UNA SOLUCIÓN VÁLIDA?

Las Directrices aclaran que el artículo 50.2 contiene **dos elementos acumulativos**. No basta con cumplir uno de ellos: se debe incorporar una marca legible por máquina y se debe proporcionar un medio de detección que permita localizar la marca. Además, las directrices establecen una serie de cualidades que deben existir a lo largo del proceso.

5.1. La marca legible por máquina

Una marca es legible por máquina cuando un programa puede reconocerla o extraerla sin que una persona tenga que interpretarla visualmente. Puede insertarse durante la generación, en la fase de inferencia o mediante un proceso posterior.

El Reglamento mantiene neutralidad tecnológica. Entre las técnicas posibles se encuentran:

- **Metadatos de procedencia:** información adjunta al archivo sobre su origen o transformación.
- **Metadatos firmados criptográficamente:** permiten comprobar su integridad y detectar alteraciones.
- **Marcas de agua imperceptibles:** patrones integrados en los píxeles, muestras de audio, fotogramas o selección de palabras.
- **Huellas digitales:** representaciones técnicas que ayudan a reconocer posteriormente un contenido.

- **Registros:** permiten contrastar si un resultado fue producido por un sistema determinado.

No todas ofrecen la misma resistencia ni sirven para cualquier modalidad. Los metadatos pueden ser ricos e interoperables, pero tienden a perderse al copiar o transformar el archivo. Las marcas de agua viajan dentro del contenido, aunque pueden degradarse o ser atacadas. Los registros pueden resultar útiles para verificar un resultado concreto, pero plantean problemas de escalabilidad, privacidad y dependencia del proveedor.

Las Directrices no convierten el artículo 50.2 en una obligación general de documentar toda la procedencia creativa. El marcado debe permitir reconocer el origen artificial, pero no tiene que reconstruir necesariamente cada paso del proceso, identificar al usuario o resolver quién es el autor jurídico de la obra.

5.2. El medio de detección

El proveedor debe **poner a disposición una solución que permita localizar e interpretar la señal**. Puede consistir en:

- Una especificación pública que permita a terceros construir sus propios detectores.
- Una herramienta local o biblioteca de software.
- Un servicio en la nube.
- Una API integrada en plataformas y flujos de verificación.
- Un detector compartido entre varios proveedores.

La herramienta debe encontrarse **disponible para las personas y actores que puedan necesitarla**. Las Directrices mencionan a usuarios, personas expuestas al contenido, autoridades, investigadores, medios, verificadores, organizaciones de la sociedad civil, plataformas y titulares de derechos. El acceso puede adaptarse al contexto. Si el contenido solo circula en un entorno profesional cerrado, con una audiencia limitada y salvaguardas eficaces frente a su difusión externa, el detector puede restringirse a esa audiencia.

Además, es importante la **interoperabilidad** de la solución. Las Directrices señalan que, cuando existan, los proveedores deben recurrir a soluciones de detección de acceso público basadas en normas del sector que permitan a terceros implementar sistemas de detección. Durante la fase inicial de aplicación del artículo 50.2, el proveedor puede utilizar una solución propia, de terceros o compartida, siempre que ofrezca una interoperabilidad suficiente; esta flexibilidad es transitoria y no elimina el objetivo de avanzar hacia mecanismos comunes.

En septiembre de 2026 ya existen referencias relevantes para la **capa de procedencia mediante metadatos**: la [especificación técnica abierta C2PA](#) permite incorporar y validar credenciales firmadas, y la [ISO/IEC 21617-1:2026 — JPEG Trust](#), publicada en agosto de 2026, establece un marco de confianza para la autenticidad, procedencia, atribución, propiedad intelectual e integridad de activos multimedia. Sin embargo, conviene precisar su alcance: **no son estándares armonizados de la Unión que generen por sí solos una presunción de conformidad con el artículo 50.2**, ni proporcionan un detector universal e independiente del proveedor para todas las marcas de agua. La interoperabilidad se encuentra más desarrollada en los metadatos firmados; la detección de marcas de agua continúa fragmentada entre tecnologías y claves de distintos proveedores.

5.3. El resultado debe poder entenderse

El detector no debería limitarse a mostrar una puntuación opaca. Conforme al artículo 50.5 y a la interpretación de las Directrices, **el resultado debe presentarse de forma clara, distinguible y accesible** cuando una persona decide verificar el contenido.

El Código de Buenas Prácticas propone que se indique, en la medida de lo posible:

- Si la detección se basa en metadatos, una marca de agua, una técnica forense u otro método.
- Qué proveedor o sistema ha incorporado la señal, cuando esa información está disponible.
- Qué parte del contenido parece generada o manipulada.
- Dónde se localiza la intervención artificial.
- Qué nivel de confianza y qué limitaciones presenta el resultado.

Esta última explicación es esencial. "No se ha detectado una marca" no significa "este contenido es humano". Puede tratarse de un sistema no compatible, una marca eliminada, un contenido anterior a la aplicación de la obligación o una transformación que haya degradado la señal.

Mi criterio: el resultado negativo del detector no debe formularse como una prueba de origen humano. Los apartados 69 a 78 de las Directrices articulan la detección alrededor de la marca incorporada por un sistema. Por tanto, el resultado más prudente no es "contenido no generado por IA", sino "no se ha detectado una marca compatible con las técnicas analizadas". Esta diferencia debería incorporarse tanto a la interfaz como a la documentación y a cualquier informe descargable.

5.4. Las cuatro cualidades de una solución válida

El artículo 50.2 no acepta cualquier marca o detector. Exige que las soluciones sean efectivas, fiables, robustas e interoperables, hasta donde resulte técnicamente viable.

5.4.1. Efectividad

La solución es efectiva cuando permite cumplir la finalidad del precepto: **que una persona pueda distinguir el contenido generado o manipulado por el sistema y comprender el significado del resultado de detección**. No es una propiedad exclusiva del algoritmo. Una marca puede detectarse con gran precisión y el sistema seguir siendo inefectivo si nadie encuentra la herramienta, la interfaz resulta incomprensible o la respuesta no explica qué se ha detectado.

5.4.2. Fiabilidad

La fiabilidad se refiere a la **capacidad para identificar correctamente el origen en condiciones normales y a través de contenidos de distinta extensión, tamaño, diversidad y semántica**. El proveedor debe medir errores sobre muestras propias marcadas y sobre contenidos que no llevan su marca. Los falsos positivos son especialmente delicados: atribuir a una IA un texto humano puede perjudicar a un estudiante, un periodista o un autor. Los falsos negativos también debilitan la finalidad de transparencia, pero ambos errores deben comunicarse de forma separada y comprensible.

5.4.3. Robustez

La robustez **mide si la marca y el detector mantienen su rendimiento después de transformaciones habituales o ataques deliberados**. El Código menciona, entre otras pruebas:

- Compresión, filtros, ruido y cambios de formato.
- Capturas de pantalla o grabaciones de pantalla.
- Recortes, giros, escalado, cambios de velocidad o tono.
- Paráfrasis, sustituciones léxicas y ciclos de traducción.
- Impresión y posterior escaneado, o reproducción y nueva grabación.
- Eliminación, copia, regeneración o falsificación de la propia marca.

Un ejemplo de un posible problema de robustez es el *analog hole*: un usuario puede reproducir un vídeo en una pantalla y grabarlo de nuevo con una cámara. El contenido sigue siendo perceptiblemente el mismo, pero su representación digital ha cambiado. Una solución robusta debe evaluar hasta qué punto sobrevive a este recorrido.

5.4.4. Interoperabilidad

La interoperabilidad pretende **evitar que cada proveedor cree un ecosistema cerrado en el que solo su propia herramienta pueda reconocer sus marcas**. La detección debe funcionar, hasta donde resulte posible, entre distintos sistemas, actores, contextos e implementaciones.

El Código reconoce que, en el momento de su publicación, la interoperabilidad se encontraba desarrollada principalmente para los metadatos firmados. Por eso adopta una **implantación progresiva** y pide a sus signatarios que, antes del 2 de febrero de 2027, incorporen al menos una solución para encaminar la verificación hacia la interoperabilidad deseada mediante la inclusión de: una interfaz estándar, una señal pública que indique qué detector utilizar, una herramienta compartida entre proveedores o una solución equivalente.

Mi criterio: un detector propio no satisface por sí solo la idea material de interoperabilidad. Puede ser una solución transitoria, pero la persona no debería necesitar conocer de antemano qué empresa pudo generar el contenido para saber qué herramienta utilizar. Mientras no exista un detector verdaderamente transversal, considero imprescindible que la marca contenga o permita obtener una señal de encaminamiento pública y que el proveedor documente cómo puede integrarse la verificación en servicios de terceros.

5.4.5. Viabilidad técnica, costes y estado de la técnica

Estos cuatro requisitos **no son absolutos**. El Reglamento ordena aplicarlos **en la medida en que sea técnicamente viable** y permite considerar las limitaciones de cada modalidad, los costes y el estado de la técnica generalmente reconocido.

La viabilidad técnica no depende exclusivamente de los recursos de una empresa concreta. Que un proveedor no haya invertido en una solución disponible no convierte esa solución en inviable. Al mismo tiempo, el Reglamento no exige adoptar una tecnología experimental que todavía no ofrece un rendimiento razonable.

El estado de la técnica tampoco es una fotografía permanente. Las marcas y detectores deben reevaluarse durante el ciclo de vida del sistema. Una solución que resulte proporcionada en 2026 puede dejar de serlo cuando aparezcan estándares maduros,

ataques nuevos o métodos claramente superiores. A fecha de septiembre de 2026, varios proveedores están adoptando una arquitectura híbrida: combinan metadatos firmados basados en estándares abiertos como C2PA con marcas de agua y detectores propios. [Adobe](#), [OpenAI](#) y [Microsoft](#) ya incorporan Content Credentials en determinados contenidos; Google combina C2PA con SynthID; y Anthropic se encuentra en una fase de implantación parcialmente operativa: ha documentado una adaptación de SynthID-Text y ya ofrece su API de detección en vista previa privada, pero no ha confirmado el despliegue general del marcado en los modelos actuales de Claude. Las credenciales C2PA para archivos compatibles también se presentan como una función anunciada. Existe, por tanto, un grado relevante de interoperabilidad en la capa de metadatos, que puede ser interpretada por herramientas conformes de distintos actores, pero la detección de marcas de agua continúa fragmentada y depende en gran medida del proveedor que haya generado el contenido.

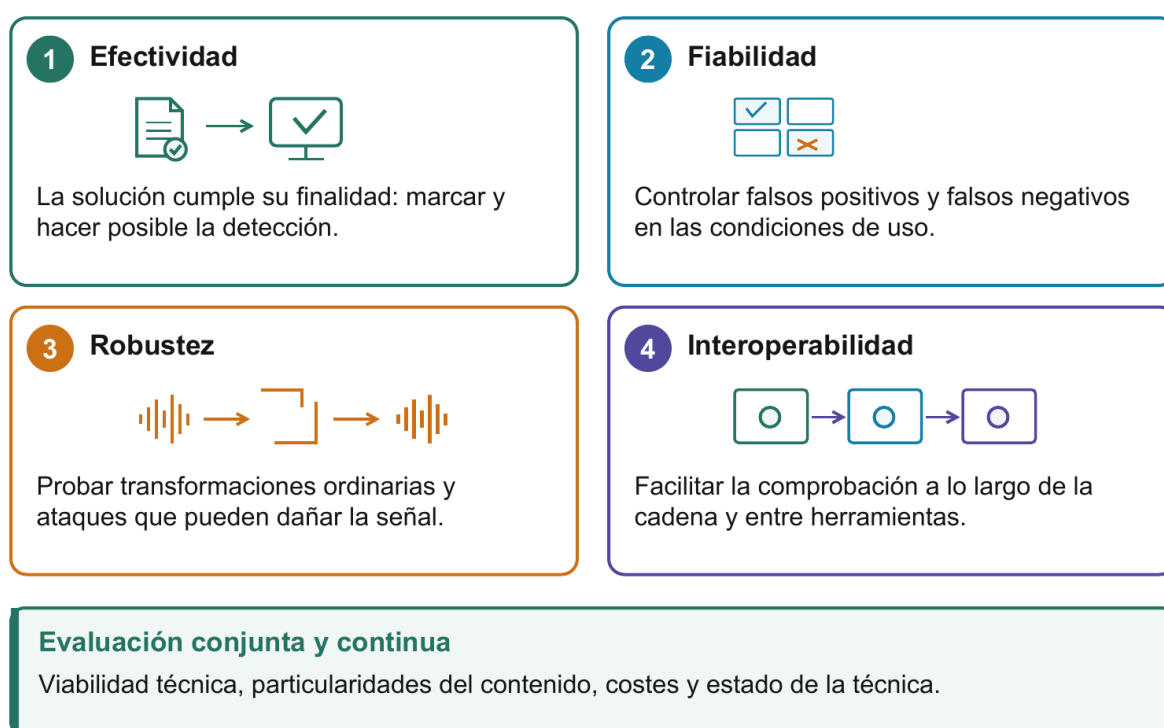


FIGURA 3. Esquema práctico para diseñar, probar, documentar, monitorizar y actualizar una solución de marcado y detección conforme al artículo 50.2.

6. CÓMO TRADUCE EL CÓDIGO LA OBLIGACIÓN A LA PRÁCTICA

Las Directrices responden principalmente a la pregunta "**¿cuándo se aplica el artículo 50.2?**". El Código de buenas prácticas trata de responder a la siguiente: "**¿cómo puede aplicarlo un proveedor con las tecnologías disponibles?**". La adhesión es voluntaria. Para sus signatarios, sin embargo, los compromisos formulados como *will* representan las medidas que se comprometen a cumplir; las recomendaciones identificadas como *encouraged* y las posibilidades formuladas como *may* tienen otro grado de exigencia dentro del propio instrumento. La Comisión y el Consejo de IA han considerado adecuado el Código. Seguirlo puede facilitar la demostración del cumplimiento, pero no produce una

inmunidad automática. La autoridad de vigilancia podrá comprobar si la solución funciona realmente y si el proveedor satisface el artículo 50.2 en su contexto concreto.

La sección dedicada a los proveedores se organiza en cuatro compromisos:

1. Marcado del contenido generado o manipulado mediante IA.
2. Detección de las marcas.
3. Calidad de las soluciones de marcado y detección.
4. Pruebas, verificación, seguimiento y demostración del cumplimiento.

A continuación, examinaremos una serie de elementos del Código que profundizan lo que hemos recogido de las directrices y que ayudan a los proveedores a establecer sus procesos de marcado.

6.1. Una arquitectura de marcado por capas

Mientras no exista una técnica única capaz de satisfacer por sí sola todos los requisitos, el Código propone un enfoque de varias capas. Para audio, imagen, vídeo y texto contenido en archivos estructurados, la referencia combina:

1. **Metadatos firmados digitalmente y, cuando sea posible, sellados temporalmente.**
2. **Una marca de agua imperceptible integrada en el contenido.**

La lógica es complementaria. Los metadatos pueden proporcionar información precisa y verificable sobre la procedencia. La marca de agua pretende sobrevivir cuando esos metadatos se eliminan durante una transformación ordinaria.

Tabla 3. Arquitectura de marcado por capas propuesta por el Código de buenas prácticas

Modalidad o contexto	Solución de referencia del Código	Observación principal
Imagen, audio y vídeo	Metadatos firmados + marca de agua imperceptible	Las dos capas deben evaluarse conjuntamente
Texto contenido en PDF, Word, HTML u otro archivo estructurado	Metadatos firmados + marca de agua imperceptible	El archivo permite transportar metadatos
Texto libre	Marca de agua imperceptible	No puede confiarse en metadatos asociados a un contenedor estable
Texto libre de menos de 200 <i>tokens</i>	Excepción temporal al <i>watermarking</i> en el Código	Se considera demasiado breve para alcanzar una fiabilidad básica con el estado de la técnica actual
Producto físico en un entorno técnico cerrado	Puede bastar una sola capa en condiciones estrictas	Deben existir medidas eficaces para impedir que el resultado salga del entorno

Mi criterio: los 200 *tokens* no constituyen una excepción jurídica autónoma. El artículo 50.2 no contiene ese umbral. El Código lo utiliza para concretar qué considera técnicamente viable con las marcas de agua de texto disponibles. Si las soluciones mejoran, el concepto de "texto muy breve" debe revisarse. Un proveedor no signatario no puede limitarse a invocar la cifra sin analizar el estado de la técnica y su propia solución; y un signatario deberá revisar el umbral cuando aparezcan métodos capaces de marcar textos más breves con una fiabilidad suficiente.

El Código permite utilizar huellas digitales o registros como medidas complementarias, pero no los considera suficientes por sí solos. Un registro puede comprobar que un archivo fue generado por el sistema, pero exige consultar una base de datos y puede no funcionar cuando el contenido ha cambiado. Además, no debe interpretarse esta medida como una obligación general de almacenar las instrucciones del usuario, monitorizar sus conversaciones o conservar todos los resultados.

6.2. Preservar las marcas durante el recorrido

Insertar la señal en el momento de la generación es solo el principio. El contenido puede atravesar editores, gestores de contenidos, plataformas, redes sociales y aplicaciones de mensajería antes de llegar a su audiencia.

El Código pide a sus signatarios que realicen esfuerzos para **conservar los metadatos existentes cuando transforman contenido de entrada y que no los eliminen o alteren intencionadamente, salvo que exista una razón legítima**. También propone incluir en las condiciones o documentación una prohibición de manipular deliberadamente las marcas y evitar la comercialización de herramientas cuya finalidad sea eludirlas.

Esta medida refleja una realidad: el proveedor original no puede garantizar el comportamiento de todos los actores posteriores. Sí puede diseñar su producto para no destruir las señales que recibe, documentar sus dependencias y dificultar la eliminación maliciosa.

6.3. Información completa, no limitarse a lo esencial

Como medida recomendada, el Código anima a incorporar datos adicionales en los metadatos:

- Nombre del sistema y del proveedor.
- Momento de la generación o manipulación.
- Identificador y versión del modelo.
- Operaciones realizadas, como la eliminación de un objeto.

El Código recomienda y anima a los proveedores de sistemas de IA a incorporar información complementaria en los metadatos. Esta información puede **mejorar la trazabilidad y reducir el riesgo de que una señal genérica se interprete como una afirmación sobre la autoría**, cuestión que el marcado no resuelve. Sin embargo, su incorporación no es obligatoria. Lo más probable es que gran parte de esta información no se recoja en todos los métodos de marcado.

Además, si se incluye esta información, debe **limitarse a lo necesario y respetar las disposiciones del RGPD, incluido el principio de minimización**. No resulta preciso identificar al usuario, insertar su nombre o registrar la instrucción utilizada para cumplir el artículo 50.2. La procedencia técnica no justifica crear un sistema de vigilancia de la actividad creativa.

El Código también anima a ofrecer una **función opcional para que el usuario añada una etiqueta visible cuando necesite cumplir el artículo 50.4**. Esta ayuda puede ser muy útil, pero no traslada al proveedor la responsabilidad del responsable del despliegue: quien publica una ultrafalsificación o un texto cubierto por el artículo 50.4 sigue teniendo que comprobar que la divulgación perceptible se realiza de forma adecuada.

6.4. Un detector para cada técnica

El Código exige a sus signatarios **ofrecer un mecanismo de detección para cada técnica de marcado utilizada**. La solución **puede ser propia, compartida o facilitada por un proveedor anterior, pero debe ser accesible para los actores que necesiten verificar el contenido**.

La regla de referencia es el **acceso gratuito**. El Código contempla una posibilidad limitada de cobrar por volumen a determinados proveedores con menos de un millón de usuarios mensuales y costes operativos sustanciales. Incluso en ese supuesto, debe preservarse el acceso gratuito sin límite de volumen para autoridades, medios, verificadores, investigadores, instituciones educativas y organizaciones de la sociedad civil enumeradas en el instrumento.

Para el público general, **el detector debe ofrecer una interfaz adecuada a personas sin conocimientos especializados**. En el caso del texto libre, el Código acepta restricciones temporales a usuarios expertos cuando una herramienta de baja fiabilidad pueda producir resultados engañosos. Esa restricción no debería convertirse en permanente: debe desaparecer cuando surjan mecanismos suficientemente fiables y robustos.

Cuando una persona solicita la verificación de un contenido concreto, el Código prevé que pueda descargar el resultado en un formato firmado, con una huella del contenido analizado, la identificación del detector y un sello temporal. Esa evidencia puede resultar relevante en procesos internos, verificaciones periódicas o actuaciones de supervisión.

6.5. Detectar sin convertir el detector en una fuente de riesgo

Un verificador puede recibir documentos confidenciales, grabaciones privadas, fotografías de menores o información personal. **Obligar a subirlos a una herramienta remota crea un tratamiento adicional** y puede revelar a un tercero precisamente el contenido cuya autenticidad se desea comprobar. El Código dedica una medida específica a esta tensión. Para los servicios que requieren cargar un archivo, sus signatarios se comprometen, entre otras cuestiones, a cumplir las siguientes disposiciones (especialmente en materia de protección de datos):

- Tratar el contenido como **información sensible desde el punto de vista de la privacidad**.
- Proteger su **confidencialidad e integridad en tránsito y en reposo**.
- **Minimización**: limitar los datos a los necesarios para la detección.
- **No reutilizar el contenido para otras finalidades** incompatibles.
- Aplicar, como regla, una **política de retención cero y eliminar el contenido inmediatamente después de la verificación**.
- Adoptar **salvaguardas de seguridad y respetar las reglas sobre transferencias internacionales**.
- Favorecer, cuando resulte posible, el análisis local o la carga de una representación que **no revele la semántica del contenido**.

La privacidad no es un elemento externo al artículo 50.2. Condiciona la forma en que la detección puede ponerse a disposición sin crear un riesgo mayor que el que pretende mitigar.

6.6. Probar, documentar y actualizar

El cuarto compromiso del Código evita que el cumplimiento se reduzca a instalar una tecnología y olvidarla. **Los signatarios deben documentar un proceso proporcionado de cumplimiento y probar las soluciones antes de colocar el sistema en el mercado o ponerlo en servicio, y de forma regular durante su ciclo de vida.**

Hasta que existan métodos y bancos de pruebas reconocidos, el Código admite referencias internas y buenas prácticas del sector. También recomienda recurrir a expertos independientes, ejercicios de *red teaming* o entornos controlados de pruebas cuando sea apropiado.

La documentación debería permitir responder, al menos, a las siguientes preguntas:

1. ¿Qué salidas genera o manipula el sistema?
2. ¿Qué técnicas se aplican a cada modalidad?
3. ¿Cómo se accede al detector y quién puede utilizarlo?
4. ¿Qué tasas de error presenta en condiciones normales?
5. ¿Qué transformaciones y ataques se han probado?
6. ¿Qué sucede cuando se actualiza el modelo o se retira el detector?
7. ¿Cómo se protegen los contenidos sometidos a verificación?
8. ¿Qué deficiencias se han identificado y qué correcciones se han aplicado?

Si un detector deja de prestarse, el Código exige a sus signatarios sustituirlo por una solución igual o mejor y compatible con el contenido marcado anteriormente. La obligación no puede funcionar si las marcas insertadas hoy se vuelven indetectables cuando el proveedor cambia de versión o abandona el servicio.

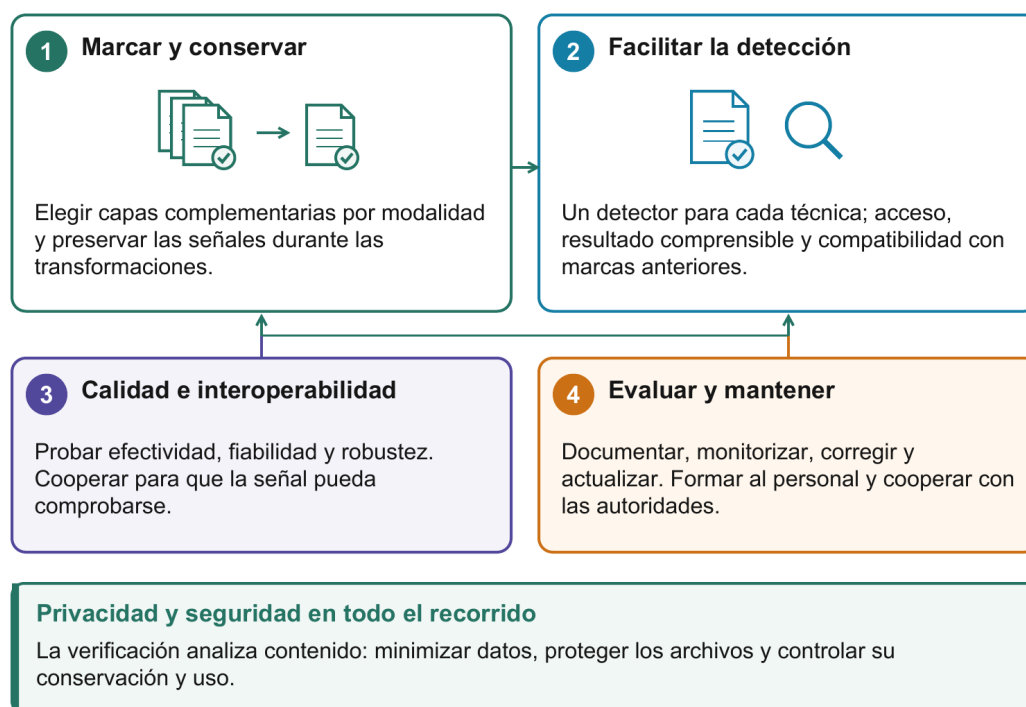


FIGURA 4. Síntesis de los cuatro compromisos y del recorrido de la señal conforme al Código de buenas prácticas.

7. ¿CÓMO ABORDAN ESTE TIPO DE MECANISMOS EMPRESAS COMO ANTHROPIC Y GOOGLE?

7.1. Anthropic y la implantación en curso del marcado de texto

El 14 de agosto de 2026, Anthropic [presentó un sistema de marca de agua para los textos generados por futuras versiones de Claude](#). Es un ejemplo especialmente útil porque el texto libre constituye una de las modalidades más difíciles de marcar.

El estado de la implantación debe describirse con precisión. La tecnología y la infraestructura de detección son, al menos, parcialmente operativas: Anthropic ya ha puesto su API de detección a disposición de determinadas organizaciones en vista previa privada. Esto permite inferir que existen modelos o muestras marcadas en un entorno controlado, de prueba o de despliegue limitado. Sin embargo, no demuestra que todas las respuestas de los modelos actuales de Claude incorporen ya la marca. La empresa continúa refiriéndose a los «futuros modelos de Claude» y anuncia que extenderá el marcado a modelos introducidos antes del 2 de agosto de 2026 durante los próximos meses, sin identificar públicamente las versiones en las que se encuentra activo. Por ello, la calificación más rigurosa es **implantación en curso y parcialmente operativa, sin despliegue general públicamente confirmado**.

La solución modifica ligeramente las probabilidades con las que el modelo selecciona unas palabras frente a otras. El resultado contiene un patrón estadístico distribuido a lo largo del texto que una herramienta especializada puede buscar después.

La marca presentada no se basa en caracteres ocultos, espacios invisibles o información añadida al archivo. Tampoco identifica a la persona, la organización o la conversación que originó el texto. Anthropic afirma que no añade *tokens* ni incrementa el coste de generación y que el sistema utiliza una versión del enfoque SynthID-Text desarrollado por Google DeepMind. La técnica modifica la fuente de aleatoriedad utilizada para elegir entre alternativas lingüísticas admisibles, pero Anthropic declara que, en sus pruebas internas, no ha observado un impacto práctico sobre el significado, la calidad, la creatividad o la legibilidad del texto. Estas afirmaciones acreditan el diseño y las pruebas internas comunicadas por la empresa, no el despliegue general del marcado en todas las versiones actuales. La señal afecta al mecanismo de selección, no pretende alterar semánticamente el contenido.

Esta arquitectura resulta bastante interesante porque, cuando se copia un texto desde la interfaz de una aplicación, se pierde el contenedor y pueden desaparecer sus metadatos. Para acompañar al texto, la marca debe encontrarse dentro de la propia distribución lingüística y no quedarse atada a una interfaz o a un documento.

La API de detección en vista previa está diseñada para evaluar la probabilidad de que Claude haya participado en la producción del fragmento. No pretende distinguir con certeza entre un texto generado completamente por el sistema y un texto humano que Claude haya editado de forma intensa, una limitación que coincide con el análisis que hemos venido realizando en este artículo. Esa precisión es jurídicamente relevante. El artículo 50.2 cubre tanto generación como manipulación, pero el resultado de la verificación no debería exagerar lo que la evidencia permite concluir. Detectar el patrón significa que el sistema probablemente intervino; no demuestra quién tuvo la idea, quién revisó el resultado o quién asume la responsabilidad editorial.

7.1.1. Los límites reconocidos

La propia empresa identifica varias situaciones de menor rendimiento:

- Fragmentos cortos.
- Pasajes muy factuales, en los que existen pocas alternativas lingüísticas.
- Código.
- Correcciones o ediciones ligeras.
- Reescrituras profundas que sustituyen por completo la formulación original.

Anthropic afirma que la marca prevista se conservará en las traducciones generadas por Claude, pero que una reescritura integral podrá eliminarla. Esta diferencia muestra una tensión inevitable: cuanto más depende la señal de los patrones del lenguaje, más expuesta se encuentra a la transformación del texto.

Desde la [actualización publicada por Anthropic el 1 de septiembre de 2026](#), la API de detección ya se encuentra en **vista previa privada**. El acceso está disponible para determinadas organizaciones elegibles (entre ellas autoridades, fuerzas de seguridad, medios, verificadores, investigadores independientes, instituciones educativas y organizaciones de la sociedad civil de la Unión) y para empresas obligadas a verificar el marcado a efectos de su propio cumplimiento. Anthropic permite registrar interés y anuncia una ampliación progresiva, pero no ofrece todavía un detector de acceso general. Por tanto, la existencia de la API ya está documentada; su disponibilidad pública y su rendimiento externo siguen siendo limitados o no verificables para el público general.

7.1.2. Metadatos anunciados para archivos creados mediante Claude

Anthropic anuncia que, cuando Claude produzca determinados archivos compatibles (menciona expresamente los formatos `.png`, `.jpg` y `.svg`), incorporará credenciales C2PA a sus metadatos. La documentación revisada no confirma que esta función esté ya desplegada de forma general. Esta segunda vía es adecuada para un archivo estable, pero no resuelve el texto que se copia y pega sin contenedor. La combinación ilustra la propuesta del Código: no existe una técnica universal. El proveedor debe adaptar el marcado a la modalidad y proporcionar una forma coherente de interpretar cada señal.

Crítica: Anthropic ofrece un ejemplo real y suficientemente concreto de una implantación en curso y parcialmente operativa para aplicar este tipo de marcado sin construir un clasificador jurídico previo para cada resultado. La API de detección ya existe en vista previa privada, mientras que el despliegue general del marcado continúa sin confirmación pública. La empresa prevé aplicarlo globalmente desde su lanzamiento y la señal se incorporará durante la generación siempre que Claude disponga de suficientes alternativas lingüísticas. Cuando se despliegue de forma general, podrán quedar técnicamente marcados resultados que estarían excluidos de la obligación. Esto tampoco permite afirmar que cualquier corrección menor vaya a ser detectada: Anthropic explica que la marca solo puede incorporarse en las palabras elegidas por Claude y que, en una revisión ortográfica o gramatical ligera, esas intervenciones pueden ser demasiado escasas para registrar una señal detectable. El verdadero caso gris aparece en la edición intensa de un texto humano, porque el detector puede indicar una probable intervención de Claude sin distinguirla de una generación completa. Como hemos señalado, la tensión entre informar sobre la intervención artificial y preservar una percepción proporcionada del mérito humano es un aspecto que debe tenerse en cuenta. ¿Es una solución perfecta? No. ¿Es una de las medidas técnicamente más prometedoras a día de hoy para que los proveedores implementen esta obligación de transparencia?

Probablemente sí, aunque su rendimiento y despliegue general todavía no pueden verificarse externamente.

7.2. SynthID y las credenciales de contenido de Google

Google lleva varios años desarrollando [SynthID para identificar contenido generado por sus sistemas](#). La tecnología incorpora marcas de agua imperceptibles en imágenes, audio, vídeo y texto, adaptadas a la estructura de cada modalidad.

En una imagen, la señal se integra en los valores de los píxeles. En un audio, se distribuye dentro de su representación sonora. En un vídeo, puede aplicarse a los fotogramas y a la pista de audio. En el texto, vuelve a utilizarse un patrón estadístico en la selección de palabras.

Google combina estas marcas con credenciales C2PA en un número creciente de productos. C2PA permite incorporar metadatos firmados sobre la procedencia y el historial del archivo. Esta combinación vuelve a separar dos funciones:

- La **marca de agua intenta permanecer unida al contenido** después de determinadas transformaciones.
- La **credencial aporta información más rica y verificable** cuando los metadatos se conservan.

La utilidad jurídica de SynthID no se encuentra solo en insertar la señal. Google ha desarrollado un [detector capaz de analizar contenidos generados por sus herramientas](#) y ya ofrece una [función de verificación en Gemini](#) para usuarios con la sesión iniciada. A fecha de esta revisión, permite cargar una imagen, un vídeo o un archivo de audio compatible y preguntar si fue generado o editado mediante herramientas de IA de Google. En el caso del vídeo, la verificación puede indicar los tramos en los que ha localizado SynthID. Esta función no está documentada como detector público de la marca de texto de Google.

Otro aspecto positivo es la integración del detector en una interfaz conocida. Un detector técnicamente preciso, pero escondido en una documentación para desarrolladores, no ofrece la misma efectividad que una función accesible desde un producto de uso general. La verificación de Gemini, sin embargo, exige iniciar sesión y está sujeta a formatos, duración, tamaño y cuotas de uso.

Actualmente, la verificación de SynthID en Gemini solo reconoce contenidos creados o editados por herramientas de IA de Google, aunque otros proveedores hayan empezado a adoptar la tecnología. La comprobación de Content Credentials amplía el alcance a archivos de imagen, vídeo o audio generados por productos compatibles con C2PA, pero no constituye un detector universal de IA y los metadatos pueden eliminarse. Ninguna de las dos técnicas permite concluir que un contenido sin señal sea humano.

Tabla 4. Comparación de las soluciones documentadas de Anthropic y Google

Elemento	Anthropic	Google
Modalidad central del caso	Texto libre	Imagen, audio, vídeo y texto
Técnica destacada	Patrón estadístico presentado y en implantación	SynthID adaptado a cada modalidad
Capa de metadatos	Credenciales C2PA anunciadas para determinados archivos compatibles	C2PA en un número creciente de productos
Detección	API en vista previa privada para entidades elegibles; acceso general todavía no disponible	Verificación en Gemini para imágenes, vídeo y audio; comprobación de Content Credentials compatibles
Principal fortaleza	La señal está diseñada para acompañar al texto copiado	Cobertura multimodal y verificación por segmentos
Principal limitación	Despliegue general no confirmado, fragilidad ante textos breves o reescrituras profundas y acceso restringido al detector	SynthID reconoce actualmente herramientas de Google; dependencia del ecosistema compatible y pérdida posible de metadatos

7.3. Evaluación propia a partir de evidencia pública

Ambos casos muestran una evolución hacia el modelo del Código, pero no permiten afirmar por sí solos que cada producto o versión cumpla todos los requisitos del artículo 50.2. El cumplimiento exige analizar el sistema concreto, la fecha, las modalidades, el acceso real al detector, las pruebas y las condiciones de uso. Para evitar una comparación meramente descriptiva, he utilizado como criterios las medidas 1.1 a 1.4, 2.1 a 2.4, 3.1 a 3.4 y 4.1 a 4.4 del Código de Buenas Prácticas. La clasificación empleada significa:

- **Verificado en documentación pública:** existe una afirmación oficial suficientemente concreta sobre la función analizada.
- **Implantación en curso y parcialmente operativa:** existe al menos un componente funcional con acceso limitado, pero no está acreditado el despliegue general de toda la solución.
- **Anunciado:** la empresa comunica una implantación futura o en curso, pero la función o sus condiciones completas todavía no pueden comprobarse públicamente.
- **Parcialmente documentado:** existe evidencia sobre una parte del criterio, pero no sobre todas las modalidades, versiones o exigencias.
- **No documentado públicamente:** las fuentes revisadas no permiten alcanzar una conclusión. No significa que la medida no exista internamente.

Tabla 5. Evaluación propia de Anthropic y Google a partir de la evidencia pública

Criterio derivado del Código	Anthropic	Google	Valoración
Marca legible por máquina	Marca estadística en implantación; funcionamiento limitado inferido de la API y despliegue general no confirmado	SynthID documentado en imagen, audio, vídeo y texto	Implantación en curso , con operatividad parcial inferida pero sin despliegue general verificable, en Anthropic; verificado documentalmente en el ecosistema compatible de Google
Capas complementarias	Marca de agua de texto en implantación; credenciales C2PA anunciadas para archivos compatibles	SynthID y C2PA en un número creciente de productos	En implantación para el texto y anunciado para C2PA en Anthropic; parcialmente documentado en Google: no puede presumirse que todas las salidas y productos utilicen siempre ambas capas
Conservación de las marcas	Se reconoce que una reescritura completa puede eliminar la marca de texto	Se documentan integraciones con C2PA y verificación en productos del ecosistema	Parcialmente documentado ; falta una visión pública completa de todos los flujos posteriores
Detector para cada técnica	API de detección de texto en vista previa privada; verificador propio de credenciales de archivos anunciado	Verificación de SynthID y Content Credentials integrada en Gemini para imágenes, vídeo y audio	Verificado con acceso restringido para la API de Anthropic; verificado documentalmente , con alcance dependiente del producto, en Google
Acceso y claridad del resultado	Acceso documentado para entidades elegibles y empresas sujetas a verificación; no disponible para el público general	Acceso con una cuenta apta y sesión iniciada; límites de formato, tamaño, duración y cuota	Parcialmente documentado en Anthropic; verificado documentalmente para las modalidades cubiertas en Google

Criterio derivado del Código	Anthropic	Google	Valoración
Limitaciones comunicadas	Se explican muestras cortas, texto factual, código, corrección y reescritura	Se explica el alcance por productos y que la verificación se refiere a marcas de Google	Verificado documentalmente en ambos casos, sin que ello sustituya las métricas completas de rendimiento
Fiabilidad y robustez	Se mencionan pruebas internas y la investigación de SynthID-Text, pero no se publican todas las métricas exigibles para cada versión	Existen investigación y despliegues a gran escala, pero las fuentes analizadas no prueban por sí solas cada requisito para cada producto	Parcialmente documentado en ambos casos
Privacidad y seguridad	La marca no identifica a la persona, organización o conversación; no se ha auditado la información completa del tratamiento asociado a la API privada	La verificación requiere analizar el contenido y debe valorarse junto con las condiciones del servicio aplicable	Parcialmente documentado ; no se ha realizado en este artículo una auditoría completa de los tratamientos asociados a los detectores
Compatibilidad con contenido antiguo	Anthropic trabaja en incorporar el marcado a modelos introducidos antes del 2 de agosto de 2026 y anuncia un despliegue durante los meses siguientes	Las fuentes analizadas no ofrecen una respuesta completa sobre todas las generaciones y versiones anteriores	Implantación anunciada en Anthropic; no documentado públicamente de forma suficiente en Google para este criterio concreto

Esta matriz produce una conclusión más prudente que afirmar que una empresa "cumple" o "no cumple". **Google presenta una infraestructura multimodal más desplegada y Anthropic una explicación especialmente transparente sobre los límites del marcado de texto.** Sin embargo, la información pública revisada no permite verificar de manera externa todos los requisitos de calidad, privacidad, acceso y continuidad para cada producto y versión.

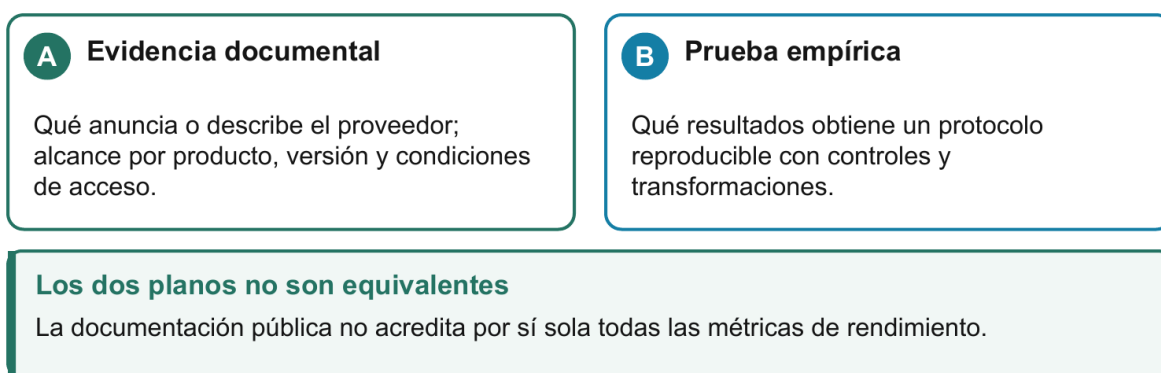
Dicho esto, ambos casos presentan elementos coherentes con el modelo del artículo 50.2 y del Código (marcado adaptado a la modalidad, mecanismos de detección y comunicación de limitaciones), pero la evidencia pública no permite afirmar que satisfagan íntegramente la obligación en todas las versiones, modalidades y condiciones de uso. La conclusión verificable es más limitada: **las soluciones están alineadas con componentes relevantes del marco y pueden facilitar la demostración del cumplimiento, que deberá evaluarse sobre el sistema concreto.**

Protocolo para una prueba empírica completa

La siguiente prueba puede utilizarse para evaluar una implementación concreta sin confundir observaciones limitadas con resultados generales:

1. Generar una muestra mediante la versión identificada del sistema y conservar el archivo o texto original.
2. Registrar la fecha, modalidad, longitud, formato y versión del producto.
3. Comprobar la muestra original con el detector oficial y conservar el resultado descargable, cuando exista.
4. Crear copias transformadas de forma controlada:
 - Para imágenes y vídeo: compresión, cambio de formato, recorte, reescalado y captura de pantalla.
 - Para audio: compresión, cambio de velocidad, ruido y nueva grabación.
 - Para texto: fragmentos de 50, 200 y 500 *tokens*, corrección ligera, paráfrasis, traducción y reescritura profunda.
5. Repetir la detección sobre cada copia y registrar el resultado, nivel de confianza, partes localizadas y mensajes de advertencia.
6. Introducir controles negativos humanos y controles generados por otros proveedores para medir falsos positivos y atribuciones incorrectas.
7. Documentar qué datos recibe el detector, durante cuánto tiempo se conservan y si existe una opción local.
8. Limitar las conclusiones a la versión, muestra y condiciones probadas.

Hasta que este protocolo se ejecute con muestras conservadas y resultados verificables, no utilizaré expresiones como "la marca sobrevive a las capturas" o "el detector reconoce correctamente las reescrituras". Precisamente esta separación entre documentación del proveedor, observación propia e inferencia es una parte esencial del análisis de cumplimiento.



Lectura de los errores de un detector

	Control negativo	Contenido marcado
Señal detectada	Falso positivo Atribuye señal artificial a un control negativo.	✓
Señal no detectada	✓	Falso negativo No reconoce la señal de un contenido marcado.

Esquema conceptual: no representa resultados medidos en esta guía.

FIGURA 5. Dos niveles de evidencia y dos tipos de error. La guía realiza una evaluación documental; este esquema no presenta tasas de rendimiento ni resultados experimentales. Véase el protocolo de prueba empírica del apartado 7.3.

8. UN MÉTODO DE CUMPLIMIENTO PARA PROVEEDORES

Tomando conjuntamente el Reglamento, las Directrices y el Código, una implantación rigurosa puede organizarse en siete pasos.

Paso 1. Clasificar el sistema y sus resultados

El proveedor debe inventariar qué modalidades genera o manipula y distinguir el contenido destinado a personas de los datos técnicos o pasos intermedios. También debe identificar las funciones de edición estándar y los usos que no alteran sustancialmente el contenido de entrada.

El análisis debe realizarse por funcionalidad. Un mismo producto puede incluir un corrector ortográfico fuera del marcado y, al mismo tiempo, un generador de resúmenes sujeto al artículo 50.2.

Paso 2. Dibujar la cadena de valor

Hay que determinar qué marcas incorpora el modelo de base, qué operaciones realiza el sistema posterior y qué ocurre al exportar el resultado. Si intervienen terceros, deben definirse responsabilidades, acceso a documentación, mantenimiento de claves, continuidad del detector y notificación de cambios.

Confiar en un proveedor anterior puede reducir la carga técnica, pero no sustituye la comprobación de que la salida final conserva una señal válida.

Paso 3. Seleccionar una solución adecuada a cada modalidad

El enfoque de referencia del Código combina metadatos firmados y marcas de agua para archivos que admiten ambas capas. Para texto libre, utiliza una marca estadística y reconoce la dificultad de los fragmentos muy breves.

La elección debe documentar las razones técnicas, las alternativas consideradas y los límites conocidos. Si el proveedor se aparta del Código, debería poder demostrar que su solución ofrece un nivel equivalente o superior respecto de los cuatro requisitos del artículo 50.2.

Paso 4. Facilitar la detección

Debe existir un detector para cada marca, con una ruta de acceso proporcionada a la audiencia potencial. La interfaz debe explicar qué señal busca, qué significa el resultado y qué no permite concluir.

Cuando el detector requiera subir contenido, deben aplicarse medidas de privacidad y seguridad. Siempre que sea viable, una herramienta local o el análisis de una huella no semántica reducirá la exposición del archivo original.

Paso 5. Probar la solución en condiciones reales

Las pruebas deben cubrir contenido de diferente longitud, tamaño, idioma, estilo y complejidad. No basta con un conjunto de ejemplos favorables preparado durante el desarrollo.

También deben simularse transformaciones previsibles: copia y pegado, compresión, recorte, captura, traducción, reescritura, combinación con contenido humano y ataques dirigidos a eliminar o falsificar la marca.

Paso 6. Comunicar las limitaciones

El proveedor debería publicar información suficiente sobre las modalidades cubiertas, el modo de acceder al detector, las tasas de error, las situaciones de menor fiabilidad y los cambios de versión. No es necesario revelar secretos empresariales o detalles que faciliten ataques, pero tampoco es aceptable presentar una conclusión probabilística como una certeza.

En especial, el resultado debe diferenciar entre:

- Se ha encontrado una marca compatible.
- No se ha encontrado una marca.
- El contenido no puede analizarse con suficiente fiabilidad.
- La señal parece dañada o manipulada.

Paso 7. Monitorizar y conservar la compatibilidad

El proveedor debe revisar el estado de la técnica, registrar deficiencias, responder a informes externos y aplicar medidas correctoras. Si cambia la marca o el detector, debe mantener una vía para verificar el contenido generado con versiones anteriores.

El resultado final no es un documento estático, sino un sistema de cumplimiento continuo: inventario, diseño, pruebas, acceso, explicación, seguimiento y corrección.

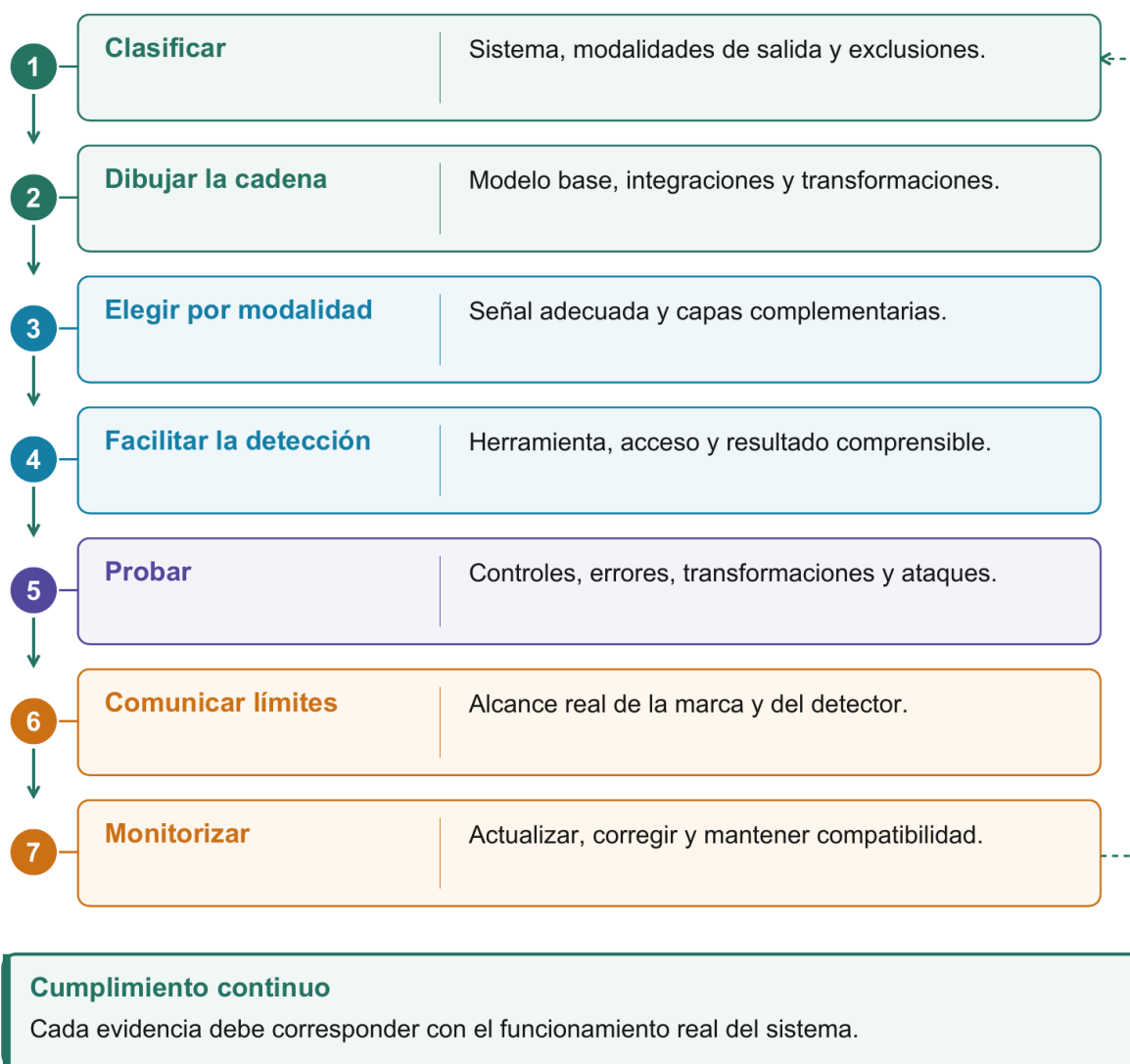


FIGURA 6. Guía visual para implantar el marcado y la detección como un sistema de cumplimiento continuo.

8.1. Matriz mínima de controles y evidencias

Los siete pasos anteriores pueden convertirse en una matriz de trabajo. Los responsables indicados son orientativos: cada organización debe asignarlos de acuerdo con su estructura y conservar evidencia de la decisión.

Tabla 6. Matriz mínima de controles y evidencias para el cumplimiento del artículo 50.2

Control	Evidencia esperada	Responsable orientativo	Riesgo si falta	Referencia principal
Inventario de salidas por modalidad y función	Catálogo de funcionalidades, formatos, destinatarios y exclusiones justificadas	Producto + Legal/Compliance	Dejar resultados sujetos sin marcar o marcar operaciones excluidas	Directrices, aps. 56-68 y 85-93
Mapa de la cadena de valor	Diagrama de modelo base, integraciones, postprocesado, exportación y terceros	Arquitectura + Producto	Perder la marca durante una transformación o asumir erróneamente que otro actor responde	Directrices, aps. 72-74; Código, medida 4.1
Marcado adaptado a cada modalidad	Especificación de metadatos, marca de agua, claves y momento de inserción	Ingeniería/ML	Solución inadecuada para el formato o fácilmente separable	Art. 50.2; Código, medida 1.1
Conservación de las señales	Pruebas del flujo de exportación, reglas de no eliminación y documentación para terceros	Ingeniería + Gestión de proveedores	Desaparición de la procedencia en usos ordinarios	Código, medida 1.2
Detector correspondiente a cada marca	Especificación, software, API o servicio y correspondencia con cada técnica	Ingeniería/ML	Marca existente pero no verificable	Directrices, aps. 69-78; Código, medida 2.1
Acceso proporcionado al detector	URL, interfaz, condiciones de acceso, cuotas y documentación	Producto + UX	Personas o verificadores incapaces de utilizar la solución	Código, medidas 2.1 y 2.4
Presentación comprensible del resultado	Diseños de interfaz, mensajes de incertidumbre, accesibilidad y resultado descargable	UX + Legal/Compliance	Convertir una inferencia limitada en una afirmación engañosa	Art. 50.5; Código, medida 2.3

Control	Evidencia esperada	Responsable orientativo	Riesgo si falta	Referencia principal
Privacidad y seguridad de la verificación	Información del tratamiento, evaluación de riesgos, retención cero, cifrado y control de accesos	Privacidad + Seguridad	Exponer documentos confidenciales o reutilizar contenido indebidamente	Directrices, ap. 94; Código, medida 2.1.3
Pruebas de fiabilidad y robustez	Banco de pruebas, controles positivos y negativos, tasas de error, transformaciones y <i>red teaming</i>	QA + ML + Seguridad	Falsos positivos, falsos negativos o eliminación sencilla de la marca	Código, medidas 3.1-3.3 y 4.2
Interoperabilidad	Estándar aplicado, señal de encaminamiento, API compartida o integración con terceros	Arquitectura + Producto	Ecosistema cerrado y verificación dependiente de conocer previamente al proveedor	Código, medida 3.4
Continuidad y compatibilidad histórica	Política de versiones, detector heredado, plan de retirada y pruebas de compatibilidad	Producto + Ingeniería	Contenido antiguo marcado que deja de ser detectable	Código, medida 2.1.4
Monitorización y corrección	Registro de incidencias, informes externos, reevaluaciones y medidas correctoras	Compliance + Producto	Mantener una solución obsoleta frente al estado de la técnica	Directrices, aps. 79-84; Código, compromiso 4
Formación del personal	Programa, contenidos, asistencia y actualización de la formación para personal técnico, de producto, cumplimiento y terceros relevantes	Compliance + RR. HH. + responsables técnicos	Aplicación inconsistente de los controles o desconocimiento de sus límites	Código, medida 4.3
Cooperación con las autoridades	Procedimiento para responder a solicitudes motivadas y facilitar documentación o acceso bajo condiciones de confidencialidad y seguridad	Legal/Compliance + Seguridad	Incapacidad para demostrar el cumplimiento o responder de forma adecuada a una investigación	Código, medida 4.4

La matriz no es una lista de documentación para archivar. Cada evidencia debe corresponder con el funcionamiento real del sistema. Una política que promete retención cero no compensa un detector que conserva los archivos; una prueba de marca en el modelo base no acredita que la aplicación final la preserve; y una interfaz accesible no corrige una tasa de falsos positivos inaceptable.

9. APLICACIÓN TEMPORAL, SUPERVISIÓN Y SANCIONES

Las obligaciones del artículo 50 son aplicables, con carácter general, desde el **2 de agosto de 2026**. Para los proveedores de sistemas comprendidos en el artículo 50.2 que ya se hubieran introducido en el mercado o puesto en servicio antes de esa fecha, el texto consolidado prevé un periodo de adaptación hasta el **2 de diciembre de 2026**. A su vez el Código de Buenas Prácticas prevé una implantación progresiva y exige a sus signatarios alcanzar antes del **2 de febrero de 2027** un nivel mínimo de interoperabilidad.

Las Directrices también aclaran que no existe una obligación retroactiva de marcar los contenidos generados o manipulados antes de que el precepto resulte aplicable. Esta precisión debe reflejarse en los detectores: la ausencia de señal en un contenido antiguo no indica incumplimiento.

El artículo 99 del AI Act incluye el incumplimiento de las obligaciones de transparencia entre las infracciones que pueden sancionarse con multas administrativas de hasta **15 millones de euros o el 3 % del volumen de negocio mundial total anual del ejercicio anterior**, con las reglas específicas del Reglamento para determinar el límite aplicable a empresas y pymes.

La sanción no se determina de forma automática por una marca perdida. Las autoridades deberán valorar la naturaleza, gravedad y duración de la infracción, la cooperación, las medidas correctoras y los demás criterios previstos. No obstante, el carácter dinámico del estado de la técnica obliga a conservar evidencia de por qué se eligió una solución y cómo se ha mantenido.

10. LOS LÍMITES DEL ARTÍCULO 50.2

La obligación ofrece una base necesaria para la transparencia técnica, pero no elimina los principales problemas de la detección de contenidos sintéticos.

- **La transparencia es invisible hasta que alguien pregunta:** La marca del artículo 50.2 puede acompañar al contenido sin que la persona sepa que existe. Su utilidad depende de que conozca el detector, pueda acceder a él y tenga un motivo para utilizarlo. El modelo resulta adecuado para investigadores, plataformas y verificadores que analizan contenidos de forma sistemática. Es menos eficaz para una persona que consume cientos de publicaciones y no comprueba cada imagen. Por eso el artículo 50.2 necesita complementarse con las divulgaciones visibles del artículo 50.4, con las medidas de las plataformas y con alfabetización mediática.
- **Un ecosistema fragmentado:** Si cada proveedor utiliza una señal y un detector propios, la persona tendría que conocer previamente qué sistema pudo generar el contenido. Esta paradoja debilita el mecanismo: para elegir el detector correcto tendría que saber parte de lo que intenta averiguar. El Código responde mediante estándares, señales de encaminamiento y detectores compartidos, pero reconoce que la interoperabilidad todavía no se encuentra plenamente madura. El plazo del 2 de febrero de 2027 para una

solución inicial de interoperabilidad muestra que el cumplimiento empieza mientras la infraestructura común sigue construyéndose.

- **Las marcas pueden degradarse y falsificarse:** Copiar un texto, comprimir una imagen o grabar una pantalla puede dañar la señal. Un actor malicioso puede intentar eliminarla. También puede copiar una marca sobre contenido que no procede del sistema para crear una atribución falsa. La robustez reduce estos riesgos, pero no los elimina. Por eso un resultado de detección debería comunicar el grado de confianza y la integridad de la señal, no limitarse a una respuesta binaria.
- **El texto sigue siendo el caso más frágil:** El texto no conserva una estructura digital estable cuando se copia. Las marcas estadísticas necesitan una extensión suficiente y pueden desaparecer con una reescritura. Además, determinados textos factuales o técnicos ofrecen poca libertad para variar la elección de palabras. Restringir temporalmente el detector a expertos puede evitar acusaciones erróneas, pero reduce la accesibilidad para el público. Abrirlo con una precisión insuficiente puede generar el problema contrario: que una herramienta aparente autoridad y atribuya falsamente un texto humano a una IA.
- **El riesgo de confundir origen, autoría y responsabilidad:** Una etiqueta técnica puede favorecer la idea simplista de que el contenido pertenece a la máquina o carece de autor. Esa conclusión no se desprende del Reglamento. La creación puede combinar decisiones humanas, fuentes previas, edición y resultados artificiales en grados muy diferentes. La mejor implementación debería describir la señal con precisión: "este sistema intervino en la generación o manipulación". No debería afirmar "está generado por IA" ni resolver cuestiones de propiedad intelectual que requieren otro análisis.
- **La detección también trata datos:** Una infraestructura global de verificación puede convertirse en un punto de concentración de documentos y comunicaciones sensibles. La política de retención cero propuesta por el Código, el procesamiento local y la minimización no son detalles accesorios: son condiciones para que la transparencia no se consiga a costa de la privacidad y la confidencialidad.
- **La ausencia de marca seguirá siendo ambigua:** Durante años convivirán contenidos humanos, resultados de sistemas no sometidos al AI Act, archivos anteriores a agosto de 2026, marcas eliminadas y sistemas que incumplen. La ausencia de una señal nunca debería presentarse como un certificado de origen humano. Este es probablemente el límite más importante del modelo. El artículo 50.2 puede aportar una prueba positiva cuando encuentra una marca válida; su capacidad para producir una prueba negativa es mucho menor.



FIGURA 7. La ausencia de marca admite explicaciones diferentes. No permite certificar autoría humana ni resolver por sí sola la atribución de responsabilidades. Síntesis del apartado 10.

11. CONCLUSIONES: UNA MARCA SOLO ES ÚTIL SI PUEDE SEGUIRSE HASTA UNA EXPLICACIÓN

El artículo 50.2 introduce una obligación técnica con una finalidad social: **permitir que las personas y los intermediarios reconozcan el contenido generado o manipulado mediante IA mediante una señal legible por máquina y un medio de detección.**

Su aplicación no consiste en insertar cualquier metadato ni en publicar un detector genérico. Exige una cadena completa:

1. Identificar qué resultados quedan dentro del ámbito del precepto.
2. Distinguir la generación o manipulación sustancial de la edición estándar.
3. Elegir una técnica adaptada a cada modalidad.
4. Combinar capas cuando una sola no ofrece suficiente calidad.
5. Poner a disposición un detector para cada marca.
6. Explicar con claridad el significado y los límites del resultado.
7. Proteger la privacidad del contenido sometido a verificación.
8. Probar la fiabilidad, robustez e interoperabilidad en condiciones reales.
9. Mantener la compatibilidad y actualizar la solución durante el ciclo de vida.

La mayor aportación del precepto no es prometer que podremos reconocer todo contenido artificial. Es establecer que los proveedores no pueden desentenderse de la procedencia una vez generado el resultado. Deben diseñar una vía técnica para que esa procedencia pueda comprobarse.

No es un sistema perfecto. La implantación de estos mecanismos puede generar tensiones con otros intereses y derechos. Si se pretende instaurar métodos de detección sostenibles, la percepción del mérito humano puede verse afectada por la inclusión de señales genéricas en contenidos que no están expresamente sujetos a marcado. De igual manera, el análisis de contenidos que pueden incluir datos personales (incluidas categorías especiales) supone un reto para los proveedores de detectores, que deberán cumplir las obligaciones del RGPD respecto de los tratamientos que realicen.

El tiempo dirá si el marcado de estos contenidos genera cargas excesivas para los proveedores, efectos reputacionales desproporcionados sobre las personas que utilizan IA o nuevos riesgos para la protección de datos y la privacidad. También permitirá comprobar si, por el contrario, cumple eficazmente su finalidad de informar a usuarios e intermediarios sobre la procedencia del contenido y mantiene esas cargas en un plano proporcionado.

REFERENCIAS

Selección de normativa, orientaciones institucionales y documentación técnica utilizada para preparar este análisis.

Última comprobación: 2 de septiembre de 2026.

- [1] [Reglamento \(UE\) 2024/1689 por el que se establecen normas armonizadas en materia de inteligencia artificial — texto consolidado](#). Marco normativo principal; véanse especialmente los artículos 3, 50, 99 y los considerandos 133 a 136.

<https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A02024R1689-20260727>

- [2] [Reglamento \(UE\) 2026/1744 — Digital Omnibus sobre IA](#). Introduce el periodo transitorio hasta el 2 de diciembre de 2026 para determinados sistemas ya comercializados.

<https://eur-lex.europa.eu/eli/reg/2026/1744/oj?locale=es>

- [3] [Comisión Europea: Directrices sobre las obligaciones de transparencia del artículo 50](#). Interpretación institucional sobre el alcance, las excepciones y las condiciones del marcado y la detección.

<https://ec.europa.eu/newsroom/dae/redirection/document/131215>

- [4] [Comisión Europea: página oficial de las Directrices del artículo 50](#). Página institucional de referencia y acceso al documento de Directrices.

<https://digital-strategy.ec.europa.eu/en/library/guidelines-transparency-obligations-providers-and-deployers-ai-systems>

- [5] [Comisión Europea: preguntas frecuentes sobre las obligaciones de transparencia del artículo 50](#). Aclaraciones institucionales actualizadas el 24 de julio de 2026.

<https://digital-strategy.ec.europa.eu/en/faqs/transparency-obligations-under-article-50-ai-act>

- [6] Código de buenas prácticas sobre el marcado y etiquetado de contenidos generados mediante IA. Arquitectura operativa de marcado, detección, pruebas e interoperabilidad.
<https://ec.europa.eu/newsroom/dae/redirection/document/129555>
- [7] Comisión Europea: página oficial del Código de buenas prácticas. Información institucional sobre el Código y su aplicación voluntaria.
<https://digital-strategy.ec.europa.eu/en/policies/code-practice-ai-generated-content>
- [8] Anthropic: Claude text watermark. Documentación de 14 de agosto de 2026, actualizada el 1 de septiembre de 2026, sobre la marca de texto y la API de detección.
<https://www.anthropic.com/news/claude-text-watermark>
- [9] Google: Our latest efforts to identify AI-generated media. Actualización de 19 de mayo de 2026 sobre SynthID, detección y credenciales de contenido.
<https://blog.google/innovation-and-ai/products/identifying-ai-generated-media-online/>
- [10] Google: How SynthID Detector helps identify content made with Google's AI tools. Descripción del funcionamiento y alcance del detector SynthID.
<https://blog.google/innovation-and-ai/products/google-synthid-ai-content-detector/>
- [11] Ayuda de Gemini: verificar imágenes, vídeos y audio generados por IA. Condiciones y límites documentados para la verificación de contenidos en Gemini.
<https://support.google.com/gemini/answer/16722517?hl=es-ES>
- [12] Google DeepMind: Scalable watermarking for identifying large language model outputs. Artículo científico sobre SynthID-Text publicado en Nature en 2024.
<https://www.nature.com/articles/s41586-024-08025-4>
- [13] C2PA: especificación técnica sobre Content Credentials. Especificación abierta para credenciales firmadas de procedencia y autenticidad.
<https://spec.c2pa.org/specifications/>
- [14] C2PA: Conformance Explorer. Listado de productos conformes y listas de confianza del ecosistema C2PA.
<https://spec.c2pa.org/conformance-explorer/>
- [15] ISO/IEC 21617-1:2026 — JPEG Trust, Part 1: Core foundation. Marco de confianza para autenticidad, procedencia, atribución e integridad de activos multimedia.
<https://www.iso.org/standard/91405.html>
- [16] Adobe: Content Credentials overview. Documentación sobre la implementación de credenciales de contenido en Adobe Firefly.
<https://helpx.adobe.com/firefly/web/get-started/learn-the-basics/content-credentials-overview.html>
- [17] OpenAI: C2PA in ChatGPT Images. Documentación sobre el uso de metadatos C2PA en imágenes generadas con ChatGPT.
<https://help.openai.com/en/articles/8912793-c2pa-in-chatgpt-images>
- [18] Microsoft 365 Copilot Chat: preguntas frecuentes sobre Content Credentials. Información sobre credenciales de contenido en determinados resultados de Microsoft 365 Copilot.
<https://support.microsoft.com/en-US/Microsoft-365-Copilot/frequently-asked-questions-about-microsoft-365-copilot-chat>

- [19] [Meta: Labeling AI-Generated Images on Facebook, Instagram and Threads](#). Descripción del uso de metadatos y declaraciones del usuario para etiquetar contenidos en las plataformas de Meta.

<https://about.fb.com/news/2024/02/labeling-ai-generated-images-on-facebook-instagram-and-threads/>