

Detección y prevención de deepfakes generados por la inteligencia artificial en los entornos de Smart Cities y Salud

Noviembre de 2025

Versión 1.0



Junta de Andalucía

Objeto del Expediente:

**“ELABORACIÓN DE GUÍAS DE CIBERSEGURIDAD IA PARA ENTORNOS DE SMART CITIES Y SALUD”
(CONTR 2024 829535).**

Esta guía forma parte de la colección Guías de Ciberseguridad en IA para las Smart Cities y el sector Salud creada por la Agencia Digital de Andalucía como parte del Proyecto Red Argos, perteneciente al programa RETECH, de Redes Inteligentes de Especialización Tecnológica, en el marco del Plan de Recuperación, Transformación y Resiliencia, financiado por la Unión Europea-NextGenerationEU, a través de INCIBE.

Autor del documento: Agencia Digital de Andalucía

Tipo de documento: Guía

Fecha de elaboración: 11/11/2025

Fecha de última actualización: 9/12/2025

ÍNDICE DE CONTENIDOS

1.	Introducción.....	6
2.	Objetivo y alcance.....	7
2.1.	Objetivo	7
2.2.	Alcance.....	7
3.	Referencias normativas	9
4.	Definiciones	12
5.	Tipologías de deepfakes y tecnologías de generación	15
5.1.	Tecnologías habilitadoras de la generación de deepfakes	15
5.1.1.	Redes Generativas Adversarias	15
5.1.2.	Modelos de Difusión.....	17
5.1.3.	Síntesis y Clonación de Voz	17
5.1.4.	Transferencia de movimiento y expresiones	18
5.1.5.	Modelos multimodales avanzados.....	19
5.1.6.	Modelos autoregresivos basados en arquitecturas Transformer	19
5.2.	Tipologías de deepfakes según el contenido manipulado	20
5.2.1.	Deepfakes de imagen.....	20
5.2.2.	Deepfakes de audio.....	21
5.2.3.	Deepfakes de vídeo	21
5.2.4.	Deepfakes multimodales.....	21
6.	Riesgos y escenarios de deepfakes en Salud y Smart Cities	22
6.1.	Riesgos transversales y sistémicos	22
6.2.	Riesgos en entornos de Smart Cities	23
6.3.	Riesgos en el sector Salud.....	24
7.	Principios de seguridad para la gestión de deepfakes.....	26
8.	Metodologías de detección y verificación de autenticidad de deepfakes	30
8.1.	Enfoque general y principios metodológicos	30
8.2.	Fase 1: Registro, Contextualización y Clasificación del Contenido Audiovisual.....	31
8.3.	Fase 2: Cribado inicial y evaluación perceptiva	33
8.4.	Fase 3: Análisis técnico intermedio	34
8.5.	Fase 4: Verificación avanzada y análisis especializado.....	36

8.6.	Fase 5: Conclusión, documentación y decisión operativa.....	37
8.7.	Capacidades organizativas necesarias.....	38
9.	Medidas de prevención y mitigación de deepfakes	40
9.1.	Medidas organizativas y de gobernanza	40
9.2.	Medidas técnicas en entornos de Smart Cities	41
9.3.	Medidas técnicas en el sector Salud	42
9.4.	Mitigación en comunicación institucional	43
10.	Recomendaciones para usuarios y profesionales.....	44
10.1.	Ciudadanía y usuarios de servicios públicos.....	44
10.2.	Personal técnico municipal y operadores de Smart Cities	45
10.3.	Personal sanitario	45
11.	Referencias.....	47
12.	Otras referencias de interés	49
12.1.	Lista de materias tratadas en la guía.....	49
12.2.	Lista de productos mencionados en la guía.....	49
12.3.	Lista de servicios mencionados en la guía	49

ÍNDICE DE TABLAS

Tabla 1: Definiciones empleadas.	14
---------------------------------------	----

1. Introducción

La rápida evolución de las tecnologías de inteligencia artificial generativa está impulsando avances relevantes en múltiples ámbitos de la gestión pública, incluyendo la optimización de servicios urbanos, la automatización de procesos administrativos y la mejora de la atención sanitaria. Modelos modernos como las Redes Generativas Adversarias (GAN), los sistemas de difusión o las arquitecturas multimodales permiten crear, transformar y mejorar contenido audiovisual con fines legítimos, como la simulación de escenarios, la formación profesional, la reconstrucción de imagen médica, la mejora de calidad visual o la generación de contenidos accesibles. Estas capacidades, cuando se integran de forma responsable, aportan eficiencia, precisión y nuevas oportunidades de innovación en los entornos urbano y sanitario.

No obstante, las mismas capacidades que permiten mejorar procesos pueden ser utilizadas de forma maliciosa, generando un riesgo creciente para la integridad, autenticidad y veracidad del contenido audiovisual. La facilidad con la que estas tecnologías pueden producir imágenes, audios o vídeos con apariencia realista ha favorecido la proliferación de deepfakes, es decir, contenidos sintéticos que imitan identidades, escenarios o eventos reales con un alto grado de fidelidad.

En las Smart Cities, donde la gestión de la movilidad, la seguridad ciudadana, la operación municipal y las comunicaciones oficiales dependen cada vez más de sistemas automatizados y de información audiovisual en tiempo real, la presencia de contenido manipulado puede distorsionar decisiones operativas, generar confusión en la ciudadanía o afectar procedimientos administrativos y sancionadores. En el sector Salud, la integridad de las imágenes diagnósticas, los registros clínicos audiovisuales y las comunicaciones asistenciales es esencial para garantizar diagnósticos fiables, decisiones terapéuticas seguras y una atención de calidad. Cualquier manipulación sintética puede comprometer la seguridad del paciente y la trazabilidad clínica, pilares fundamentales del sistema sanitario.

La utilización malintencionada de contenido sintético puede adoptar múltiples formas: suplantación de voz, manipulación de imágenes clínicas, alteración de vídeo de videovigilancia, generación de contenido institucional falso o construcción de escenarios visuales inexistentes. Estas amenazas se ven reforzadas cuando coinciden fallos en la verificación, debilidades operativas o ausencia de controles adecuados. El ENISA Threat Landscape for Artificial Intelligence identifica de forma explícita los deepfakes como una amenaza estratégica emergente, con impacto potencial en la resiliencia institucional y en la confianza pública.

Por ello, esta guía se desarrolla en el marco de un ecosistema donde la IA generativa aporta beneficios reales, pero exige medidas sólidas de autenticación, supervisión y gobernanza para prevenir usos indebidos. El documento se alinea con las obligaciones establecidas por el Reglamento Europeo de Inteligencia Artificial (RIA), la Directiva NIS2, el RGPD, la LOPDGDD, el Esquema Nacional de Seguridad (ENS), MDR/IVDR, eIDAS y los estándares internacionales ISO/IEC aplicables a inteligencia artificial, protección de datos y ciberseguridad. Asimismo, se integra en las políticas nacionales de digitalización y modernización del sector público, ofreciendo un marco estratégico y operativo para evaluar, prevenir y mitigar los riesgos derivados de la manipulación sintética en dichos sectores donde la autenticidad del contenido audiovisual constituye un requisito esencial para un funcionamiento seguro, ético y confiable.

2. Objetivo y alcance

2.1. Objetivo

La presente guía tiene como objetivo principal establecer un marco estratégico, técnico y operativo para la detección, verificación, prevención y gestión de deepfakes en entornos de Smart Cities y Salud. Proporciona una referencia metodológica y organizativa que permita reforzar la integridad, autenticidad y trazabilidad del contenido audiovisual utilizado en procesos críticos, con el fin de:

- Definir criterios verificables para la identificación de deepfakes en imagen, vídeo, audio y contenido multimodal.
- Potenciar las capacidades de análisis forense digital y autenticación audiovisual en organizaciones críticas.
- Establecer medidas de prevención y mitigación, adaptadas a los requisitos funcionales y regulatorios de los sectores Salud y Smart Cities.
- Reforzar la coordinación entre administraciones públicas, operadores de infraestructuras y organismos sanitarios mediante canales de alerta, protocolos de escalado y mecanismos de respuesta a incidentes relacionados con contenidos sintéticos.

La guía no pretende proporcionar instrucciones para el desarrollo de modelos generativos, sino orientar a organizaciones y profesionales en la gestión segura, responsable y conforme a normativa de los riesgos asociados a la manipulación audiovisual mediante IA.

2.2. Alcance

El alcance de esta guía abarca a todas las organizaciones y entidades que, en el ejercicio de sus funciones, utilizan, reciben, generan o gestionan contenidos audiovisuales cuyo uso pueda resultar crítico para la prestación de servicios en los entornos de Smart Cities y Salud. Esto incluye administraciones públicas de cualquier nivel territorial, operadores de infraestructuras urbanas, centros hospitalarios y asistenciales, organismos reguladores, unidades de emergencia y protección civil, plataformas digitales municipales, proveedores tecnológicos, así como equipos responsables de ciberseguridad, sistemas de información y cumplimiento normativo. También se dirige a los profesionales que intervienen en la creación, análisis, preservación o verificación de imágenes, audios y vídeos que forman parte de procedimientos urbanos, clínicos o administrativos en los que la autenticidad del contenido es un requisito esencial para la seguridad ciudadana, la continuidad operativa y la calidad asistencial.

Adicionalmente, y como público destinatario secundario, la guía incluye orientaciones prácticas para la ciudadanía y usuarios de servicios públicos, en la medida en que pueden verse expuestos a campañas de manipulación audiovisual y desempeñan un papel relevante en la detección temprana y en la no propagación de contenidos sintéticos (Capítulo 10).

Esta guía abarca de manera completa el ciclo de gestión del riesgo asociado a los deepfakes: desde los procesos de identificación temprana, verificación técnica y análisis forense hasta las medidas de prevención, mitigación y respuesta organizativa ante incidentes de manipulación sintética. Se delimita al análisis de los riesgos asociados a la generación o manipulación sintética de contenido audiovisual y

multimodal (imagen, vídeo, audio y combinaciones de estos), excluyendo el desarrollo o entrenamiento de modelos generativos. Esta guía es aplicable tanto a infraestructuras tecnológicamente avanzadas como a sistemas heredados y debe entenderse como una referencia estratégica para integrar la protección frente a contenidos sintéticos dentro de los marcos de ciberseguridad, gobernanza de la inteligencia artificial, protección de datos personales y resiliencia digital establecidos por las normativas vigentes.

3. Referencias normativas

Las referencias incluidas en este apartado conforman el marco jurídico, técnico y ético directamente aplicable a la detección y prevención de contenidos sintéticos generados mediante inteligencia artificial en los ámbitos urbano y sanitario. Se integran normas europeas, nacionales e internacionales que regulan la inteligencia artificial, la protección de datos, la ciberseguridad, la autenticidad digital y la trazabilidad de la información.

- **Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, Reglamento de Inteligencia Artificial (RIA):** Establece un marco jurídico armonizado para el desarrollo y uso de sistemas de inteligencia artificial en la Unión Europea. Introduce un enfoque basado en el riesgo y obligaciones específicas para los sistemas de alto riesgo, incluidos los empleados en sanidad, seguridad pública y gestión de infraestructuras urbanas críticas. Refuerza los principios de transparencia, trazabilidad, supervisión humana y rendición de cuentas.
- **Reglamento (UE) 2016/679, Reglamento General de Protección de Datos (RGPD):** Garantiza la protección de las personas físicas respecto al tratamiento de datos personales. Resulta plenamente aplicable al tratamiento de datos mediante IA, especialmente cuando involucra datos biométricos, de imagen o de voz.
- **Ley Orgánica 3/2018, de Protección de Datos Personales y garantía de los derechos digitales (LOPDGDD):** Desarrolla y complementa el RGPD en el ordenamiento español, incorporando derechos digitales y mecanismos de transparencia, supervisión y rendición de cuentas en el uso de sistemas basados en inteligencia artificial.
- **Directiva (UE) 2022/2555 del Parlamento Europeo y del Consejo, Directiva NIS2:** Refuerza las medidas para garantizar un nivel elevado de ciberseguridad en la Unión Europea. Se aplica a entidades esenciales e importantes, incluidas administraciones públicas, servicios sanitarios, operadores de infraestructuras críticas y proveedores tecnológicos vinculados a Smart Cities.
- **Esquema Nacional de Seguridad (ENS):** Define los principios y requisitos mínimos de seguridad aplicables a los sistemas de información del sector público español y de los proveedores tecnológicos. Resulta fundamental para la protección de la integridad, disponibilidad y autenticidad de la información procesada mediante sistemas de IA.
- **Reglamento (UE) 2017/745 relativo a los productos sanitarios (MDR):** Regula la seguridad, evaluación clínica y vigilancia de los productos sanitarios, incluidos aquellos que incorporan componentes de IA para diagnóstico o monitorización, garantizando la trazabilidad y la evaluación de conformidad.

- **Reglamento (UE) 2017/746 relativo a los productos sanitarios para diagnóstico in vitro (IVDR):** Establece los requisitos de validación, seguridad y trazabilidad de los dispositivos de diagnóstico in vitro, aplicables a los que incorporan algoritmos de inteligencia artificial.
- **Data Governance Act, Reglamento (UE) 2022/868:** Regula la gobernanza y la reutilización segura de datos en la Unión Europea, fomentando el intercambio confiable de información y la protección de los datos sensibles, incluidos los de carácter sanitario.
- **Digital Services Act, Reglamento (UE) 2022/2065:** Establece obligaciones de transparencia, trazabilidad y diligencia debida para prestadores de servicios digitales y plataformas en línea, aplicables a la gestión de contenidos sintéticos generados mediante IA.
- **Reglamento (UE) 910/2014, Reglamento eIDAS sobre identificación electrónica y servicios de confianza:** Proporciona el marco para la firma electrónica, el sello electrónico y el sellado de tiempo y los certificados digitales, que resultan fundamentales para la autenticación y la trazabilidad de los contenidos y las evidencias digitales.
- **Carta de Derechos Digitales (Gobierno de España, 2021):** Reconoce los derechos fundamentales en el entorno digital, incluyendo identidad, veracidad informativa, protección frente a la manipulación algorítmica y transparencia en el uso de tecnologías basadas en inteligencia artificial.
- **Ley 41/2002, básica reguladora de la autonomía del paciente:** Establece los derechos y obligaciones en materia de información y documentación clínica, garantizando la confidencialidad, la integridad y el consentimiento informado en los procesos asistenciales que emplean IA o contenidos sintéticos.
- **ISO/IEC 42001:2023 – Sistemas de gestión de la inteligencia artificial:** Establece los requisitos para implantar y mantener un sistema de gestión de IA que garantice un uso seguro, responsable y trazable de los algoritmos. Su aplicación refuerza la gobernanza, la transparencia y el control en sistemas generativos utilizados en entornos urbanos y sanitarios.
- **ISO/IEC 23894:2023 – Gestión del riesgo en la inteligencia artificial:** Define un marco estructurado para identificar, evaluar y mitigar los riesgos derivados del desarrollo y uso de la IA, incluyendo los relacionados con la generación de contenidos sintéticos y su impacto en la integridad de la información.
- **ISO/IEC 38507:2022 – Gobernanza de la inteligencia artificial:** Proporciona directrices para integrar la IA en los sistemas de gobernanza organizacional, asignando responsabilidades, supervisión de la alta dirección y control del riesgo en la toma de decisiones automatizada.
- **ISO/IEC 27001:2022 – Seguridad de la información, ciberseguridad y protección de la privacidad:** Establece los requisitos para gestionar la seguridad de la información en

organizaciones que operan con IA, asegurando la confidencialidad, integridad, disponibilidad y autenticidad de los datos procesados.

- **ISO 31000:2018 – Gestión del riesgo:** Ofrece principios y directrices generales para establecer un marco eficaz de gestión del riesgo organizativo, aplicable a los riesgos tecnológicos y algorítmicos en los sistemas inteligentes de Smart Cities y salud.
- **ISO 27799:2016 – Seguridad de la información en organizaciones sanitarias:** Adapta los controles de seguridad de ISO/IEC 27001 al sector salud, garantizando la protección de los datos clínicos frente a accesos indebidos, manipulaciones o usos no autorizados de la información médica digital.
- **ISO/IEC 30111:2019 – Gestión de vulnerabilidades:** Describe los procesos para la recepción, análisis y coordinación de vulnerabilidades en software y sistemas de IA, esenciales para prevenir incidentes que afecten la integridad de los modelos y de los datos.
- **ENISA AI Threat Landscape (última edición disponible):** identifica amenazas emergentes vinculadas a inteligencia artificial, incluyendo campañas de desinformación, manipulación sintética y riesgos asociados a contenido audiovisual.
- **NIST AI Risk Management Framework 1.0 (2023):** Proporciona un marco de referencia para gestionar los riesgos de la IA en todo su ciclo de vida, promoviendo la transparencia, la responsabilidad y la mitigación de sesgos en sistemas automatizados y generativos.

4. Definiciones

Este capítulo proporcionará una tabla clara y precisa de acrónimos y términos relevantes utilizados a lo largo del documento.

Acrónimos/Términos	Definición
Autenticidad del contenido	Propiedad que garantiza que un contenido audiovisual es genuino, procede de una fuente legítima y no ha sido generado ni alterado de forma no autorizada. Es un requisito esencial para su validez operativa, clínica o probatoria.
C2PA	Estándar abierto que permite registrar y verificar la procedencia y el historial de modificaciones de un contenido digital mediante metadatos protegidos. Facilita la identificación de contenido generado o alterado por IA.
Clonación de voz	Técnica de IA que permite generar audio sintético reproduciendo con alta fidelidad la voz de una persona real a partir de muestras previas. Puede utilizarse para suplantación de identidad y emisión de comunicaciones falsas en entornos críticos.
Contenido multimodal	Información generada o procesada mediante la combinación coordinada de múltiples modalidades (texto, imagen, audio, vídeo), típicamente producida por modelos multimodales avanzados.
Contenido sintético	Cualquier contenido total o parcialmente generado mediante sistemas de IA, independientemente de que su finalidad sea legítima o maliciosa. Incluye imágenes, audios, vídeos y combinaciones multimodales.
Deepfake	Contenido audiovisual o multimodal generado o manipulado mediante técnicas de IA generativa con el fin de simular de forma realista identidades, acciones, escenarios o comunicaciones que no corresponden a la realidad, comprometiendo la autenticidad del contenido.
Evidencia digital	Información en formato digital con valor probatorio en procedimientos administrativos, judiciales, clínicos o regulatorios. Requiere autenticidad, integridad y cadena de custodia verificable.
GAN (Red Generativa Adversaria)	Arquitectura de IA compuesta por un generador y un discriminador que compiten entre sí para producir contenido sintético cada vez más realista, ampliamente utilizada en la generación de deepfakes.
IA (Inteligencia Artificial)	Sistema basado en máquinas diseñado para operar con distintos niveles de autonomía, capaz de generar resultados como predicciones, contenidos, recomendaciones o decisiones que influyen en entornos físicos o digitales, conforme a la definición del Reglamento (UE) 2024/1689 (RIA).

Acrónimos/Términos	Definición
IA generativa	Subconjunto de sistemas de inteligencia artificial capaces de generar contenido nuevo (texto, imagen, audio, vídeo o multimodal) a partir de datos de entrenamiento, incluyendo modelos basados en GAN, difusión, transformers y arquitecturas multimodales.
Integridad del contenido	Garantía de que el contenido audiovisual no ha sido modificado desde su captura hasta su uso final. Incluye la protección frente a manipulaciones parciales o sintéticas que alteren su significado.
Metadatos	Información estructurada asociada a un archivo audiovisual que describe su captura, contexto y parámetros técnicos. Son esenciales para la verificación de autenticidad e integridad.
Modelo autorregresivo (Transformer)	Arquitectura de IA que genera contenido secuencial, como texto o audio, a partir del contexto previo. En los deepfakes actúa como base para generar discursos que luego se transforman en voz o vídeo sintético.
Modelo de difusión	Modelo generativo de IA que crea o modifica imágenes y vídeos mediante un proceso iterativo de reconstrucción a partir de ruido. Permite generar escenas visuales altamente coherentes, lo que dificulta la detección de manipulaciones en contextos urbanos y sanitarios.
Modelo multimodal avanzado	Sistema de IA que genera contenido sintético combinando texto, voz, imagen y vídeo de forma coordinada. Permite producir comunicaciones audiovisuales completas con apariencia institucional o profesional.
PACS / RIS	Sistemas utilizados para almacenar y gestionar imágenes médicas y datos radiológicos. Son infraestructuras críticas para la trazabilidad y validez clínica de la imagen diagnóstica.
PRNU	Técnica forense que identifica el patrón único de ruido de un sensor de imagen para verificar la procedencia física de fotografías o vídeos. Debe emplearse junto con otros métodos de verificación.
Sistema de alto riesgo	Sistema de IA clasificado como de alto riesgo conforme al Reglamento de IA por su impacto potencial en salud, seguridad pública o derechos fundamentales. Está sujeto a obligaciones reforzadas de control y gobernanza.
Smart City	Entorno urbano que integra tecnologías digitales y sistemas automatizados para gestionar servicios públicos, movilidad y seguridad. Depende de información audiovisual cuya autenticidad es crítica.
Supervisión humana significativa	Intervención humana efectiva en sistemas de IA de alto riesgo que permite comprender, revisar y corregir decisiones automatizadas. Es obligatoria en contextos que afectan a salud, seguridad pública o derechos fundamentales.

Acrónimos/Términos	Definición
Telemedicina	Prestación de servicios sanitarios mediante tecnologías digitales que incluyen comunicaciones audiovisuales. Requiere garantizar autenticidad, integridad y trazabilidad del contenido clínico.
Transferencia de expresiones o movimiento	Técnica que traslada gestos, expresiones faciales o movimientos corporales de una persona a la imagen de otra en un vídeo. Permite alterar el comportamiento aparente de individuos en grabaciones urbanas o clínicas.
Trazabilidad	Capacidad de reconstruir el origen, custodia y transformaciones de un contenido audiovisual a lo largo de su ciclo de vida. Permite sostener la confianza institucional y la validez jurídica del contenido.

TABLA 1: DEFINICIONES EMPLEADAS.

5. Tipologías de deepfakes y tecnologías de generación

La generación de deepfakes se basa en un conjunto de técnicas avanzadas de inteligencia artificial capaces de sintetizar, modificar o reconstruir contenido audiovisual con niveles de realismo que pueden superar la detección humana. La evolución y mejora progresiva de estas tecnologías ha ampliado de forma significativa las posibilidades de manipulación del contenido audiovisual. Como consecuencia, los deepfakes pueden utilizarse para simular identidades, comportamientos, escenas o comunicaciones que aparentan ser reales, aunque no lo son.

Esta capacidad representa un riesgo relevante en contextos donde la autenticidad del contenido audiovisual resulta crítica, como los entornos de Smart Cities y Salud. Estas capacidades, desarrolladas originalmente para aplicaciones legítimas como la simulación de entornos, la mejora de imagen o el apoyo al diagnóstico clínico, se han democratizado y hoy están al alcance de actores con competencias técnicas limitadas. Este cambio amplía la superficie de riesgo para administraciones públicas y organizaciones sanitarias, donde la fiabilidad del contenido audiovisual resulta esencial para la operación segura, la gestión urbana y la prestación de servicios.

El objetivo de este apartado es proporcionar un marco claro y comprensible sobre las principales tecnologías que permiten la generación de deepfakes, así como las tipologías de deepfakes según el contenido manipulado, con el fin de facilitar la identificación de riesgos y apoyar el desarrollo de medidas de detección y prevención.

5.1. Tecnologías habilitadoras de la generación de deepfakes

La comprensión de las tecnologías que permiten la generación de deepfakes es un elemento clave para poder identificar sus riesgos y definir medidas de detección y prevención eficaces. Estas tecnologías se basan en distintos enfoques de inteligencia artificial que permiten crear, modificar o reconstruir contenido audiovisual y textual con una apariencia realista.

En este apartado se describen las principales tecnologías habilitadoras de los deepfakes, explicando de forma clara su funcionamiento general, sus capacidades técnicas más relevantes y su impacto potencial en los entornos de Smart Cities y Salud. El objetivo no es detallar su implementación técnica, sino facilitar la comprensión de qué permiten hacer estas tecnologías y por qué representan un riesgo en contextos donde la autenticidad del contenido es crítica.

5.1.1. Redes Generativas Adversarias

Las Redes Generativas Adversarias son una de las tecnologías más utilizadas para la creación de deepfakes debido a su capacidad para generar imágenes, vídeos y audios falsos con una apariencia muy similar a la real. Estas tecnologías permiten crear contenido sintético que imita personas, escenas o situaciones reales de forma convincente, lo que dificulta su detección sin controles específicos.

Su funcionamiento se basa en el análisis de grandes cantidades de datos reales, como fotografías de personas, grabaciones de voz o secuencias de vídeo. A partir de estos datos, el sistema aprende cuáles son las características que hacen que una imagen, un vídeo o un audio parezcan auténticos. Entre estas

características se incluyen la forma del rostro, la iluminación, los gestos, el movimiento de los labios al hablar o el sonido de la voz.

El sistema está formado por dos elementos que trabajan de manera conjunta:

- **Generador:** es el componente que crea el contenido falso. Produce imágenes, vídeos o audios que reproducen las características observadas en los datos reales utilizados durante el aprendizaje. En el contexto de los deepfakes, el generador puede crear personas que no existen, sustituir el rostro de una persona por otro, modificar expresiones faciales o generar voces que suenen como las de una persona real.
- **Discriminador:** es el componente que revisa el contenido generado y comprueba si parece real o si presenta errores o incoherencias. Cuando detecta fallos, esta información se utiliza para mejorar el contenido generado en intentos posteriores.

Con el paso del tiempo, este proceso permite que el contenido generado mejore de forma progresiva. Las imágenes y los vídeos pueden mantener una iluminación coherente, movimientos faciales naturales y una sincronización adecuada entre imagen y sonido. Como resultado, el contenido sintético puede resultar difícil de distinguir del contenido real mediante una simple observación visual o auditiva.

Estas tecnologías permiten tanto crear contenido completamente nuevo, como personas o escenas inexistentes, como modificar contenido real, alterando identidades, gestos o voces. Esta capacidad es la base de muchos usos maliciosos de los deepfakes, especialmente en escenarios de suplantación de identidad y manipulación de pruebas audiovisuales.

Capacidades técnicas relevantes:

- Generación de identidades faciales inexistentes con una apariencia coherente, manteniendo rasgos visuales consistentes como la forma del rostro, la iluminación general y la expresión facial.
- Modificación de identidades en imágenes y secuencias de vídeo, mediante la sustitución de rostros o la alteración de expresiones faciales y movimientos asociados al habla, sin introducir discontinuidades evidentes.
- Alteración controlada de gestos y movimientos faciales, permitiendo cambiar la expresión o el comportamiento aparente de una persona en una imagen o vídeo.
- Generación de voces sintéticas que imitan las características generales de una voz real, como el tono, el ritmo y el estilo de habla, permitiendo atribuir mensajes falsos a personas concretas

Impacto sectorial:

- **Smart Cities:** estas tecnologías pueden utilizarse para alterar imágenes de cámaras de tráfico o seguridad ciudadana, modificar matrículas, simular incidentes urbanos que no han ocurrido o manipular pruebas visuales utilizadas en procedimientos administrativos o sancionadores. Asimismo, el contenido falso puede incorporarse a sistemas municipales y afectar a decisiones operativas o análisis automatizados.
- **Salud:** en el sector sanitario, permiten modificar imágenes clínicas, grabaciones asistenciales o comunicaciones audiovisuales, introduciendo cambios que pueden influir en diagnósticos o decisiones clínicas. También pueden utilizarse para generar vídeos o audios falsos atribuidos a profesionales sanitarios, lo que afecta a la trazabilidad clínica y a la seguridad del paciente.

5.1.2. Modelos de Difusión

Los modelos de difusión son tecnologías de inteligencia artificial que se utilizan para crear o modificar imágenes y vídeos con una apariencia realista. Su principal característica es que permiten generar escenas completas sin introducir errores visuales evidentes.

Estos modelos aprenden cómo se transforma una imagen cuando se degrada de forma progresiva. A partir de ese aprendizaje, el sistema es capaz de realizar el proceso inverso y reconstruir una imagen paso a paso hasta obtener un resultado coherente.

En la práctica, estos modelos permiten crear imágenes nuevas o modificar imágenes existentes. Pueden cambiar elementos concretos de una escena, como objetos, fondos o iluminación, sin que el resultado final muestre cortes visibles.

Capacidades técnicas relevantes:

- Generación de imágenes clínicas plausibles sin necesidad de datos reales de partida.
- Alteración avanzada de escenas urbanas modificando iluminación, objetos estructurales o elementos dinámicos.
- Condicionamiento preciso mediante instrucciones textuales, máscaras o mapas de profundidad.
- Manipulación de videovigilancia preservando continuidad visual, sombras, movimiento y coherencia temporal.

En el caso del vídeo, los modelos de difusión pueden mantener continuidad entre fotogramas consecutivos. Esto evita saltos visuales y movimientos poco naturales, lo que hace que los vídeos generados resulten difíciles de detectar sin análisis específico.

Impacto sectorial:

- **Smart Cities:** estos modelos pueden utilizarse para crear escenas urbanas falsas o modificar grabaciones de videovigilancia. También pueden alterar imágenes empleadas en sistemas de tráfico, seguridad o movilidad urbana.
- **Salud:** en el ámbito sanitario, permiten generar o modificar imágenes diagnósticas de forma realista. Estas alteraciones pueden influir en diagnósticos clínicos o afectar a la calidad de los datos utilizados en sistemas sanitarios.

5.1.3. Síntesis y Clonación de Voz

La síntesis y clonación de voz son tecnologías de inteligencia artificial que permiten generar audios sintéticos que reproducen con alta fidelidad la identidad vocal de una persona real. A diferencia de grabaciones manipuladas tradicionales, estos sistemas pueden producir mensajes completamente nuevos que mantienen una identidad vocal coherente y estable.

En la práctica, estos sistemas no actúan de forma aislada. El contenido del mensaje suele generarse previamente mediante herramientas de procesamiento del lenguaje natural (PLN), que producen el texto o las instrucciones que posteriormente se transforman en audio sintético. De este modo, la suplantación vocal combina generación lingüística y síntesis acústica en un único proceso.

Estos sistemas analizan grabaciones reales de voz para identificar patrones acústicos persistentes, como la entonación, el ritmo, las pausas y la pronunciación. Una vez aprendidos estos patrones, el sistema puede generar mensajes hablados distintos sin necesidad de que la persona haya pronunciado previamente ese contenido.

El riesgo principal de estas tecnologías es que rompen el vínculo entre identidad y presencia física, permitiendo atribuir órdenes, validaciones o comunicaciones críticas a una persona que nunca las ha emitido. Esto afecta directamente a entornos donde la voz se utiliza como elemento de confianza o validación operativa.

Capacidades técnicas relevantes:

- Generación de mensajes hablados atribuidos falsamente a una persona concreta.
- Mantenimiento de una identidad vocal consistente a lo largo de múltiples comunicaciones.
- Producción de audios con apariencia natural en contextos formales, operativos o urgentes.

Impacto sectorial:

- Smart Cities: permite suplantar a responsables operativos, personal de emergencias o autoridades municipales, emitiendo instrucciones falsas que pueden activar protocolos, movilizar recursos o alterar decisiones críticas.
- Salud: posibilita la emisión de indicaciones clínicas falsas, autorizaciones inexistentes o comunicaciones asistenciales fraudulentas, afectando a la seguridad del paciente y a la trazabilidad clínica.

5.1.4. Transferencia de movimiento y expresiones

La transferencia de movimiento y expresiones es una técnica que permite alterar el comportamiento visual de una persona en un vídeo, haciendo que aparente realizar gestos, expresiones faciales o movimientos corporales que no corresponden a la realidad.

Estas tecnologías analizan secuencias de movimiento reales y trasladan esa información a la imagen de otra persona, ajustando automáticamente la posición facial, la gestualidad y la sincronización con el audio. El resultado es una secuencia audiovisual continua en la que la persona suplantada parece actuar de forma natural.

El riesgo asociado a esta técnica reside en su capacidad para modificar el significado de una acción registrada, alterando la interpretación de comportamientos, intenciones o reacciones sin necesidad de alterar el contexto completo de la escena.

Capacidades técnicas relevantes:

- Alteración de expresiones faciales y microgestos relevantes para la interpretación humana.
- Modificación de movimientos de cabeza, postura y comportamiento corporal.
- Sincronización precisa entre gestualidad y audio asociado.

Impacto sectorial:

- **Smart Cities:** esta técnica permite manipular grabaciones de videovigilancia, distorsionar el comportamiento aparente de personas en incidentes urbanos o alterar pruebas utilizadas en procedimientos administrativos o sancionadores.
- **Salud:** la transferencia de movimiento puede suplantar profesionales en teleconsultas, alterar evaluaciones clínicas grabadas o manipular intervenciones asistenciales registradas. Estas alteraciones afectan la validez clínica, la trazabilidad y la confianza de pacientes y profesionales.

5.1.5. Modelos multimodales avanzados

Los modelos multimodales son sistemas de inteligencia artificial que permiten generar contenido sintético combinando imagen, audio y vídeo de forma coordinada, produciendo secuencias audiovisuales coherentes y realistas.

En estos modelos, el procesamiento del lenguaje natural (PLN) desempeña un papel habilitador. El texto generado define el discurso, la estructura del mensaje y la intención comunicativa, y actúa como elemento de control del resto de componentes. A partir de ese contenido lingüístico, el sistema genera la voz, las expresiones faciales y la gestualidad asociada.

El resultado es un vídeo sintético completo en el que lo que se dice, cómo se dice y cómo se representa visualmente están alineados. Esta coherencia reduce significativamente las señales habituales de manipulación y dificulta su detección.

Capacidades técnicas relevantes:

- Generación de vídeos completos a partir de contenido lingüístico generado automáticamente.
- Coordinación entre texto, voz, expresiones faciales y contexto visual.
- Producción de comunicaciones audiovisuales con apariencia institucional o profesional.

Impacto sectorial:

- **Smart Cities:** los modelos multimodales pueden generar ruedas de prensa falsas, alertas urbanas fraudulentas o mensajes institucionales manipulados con apariencia plenamente legítima, afectando la confianza pública, la coordinación operativa y la gestión de emergencias.
- **Salud:** permiten producir vídeos falsificados de profesionales sanitarios, consignar instrucciones clínicas inexistentes y generar documentación audiovisual que parece auténtica. Esto puede comprometer la seguridad del paciente, influir en decisiones clínicas y erosionar la confianza institucional.

5.1.6. Modelos autoregresivos basados en arquitecturas Transformer

Los modelos autorregresivos basados en arquitecturas Transformer son sistemas de inteligencia artificial que permiten generar contenido secuencial coherente, como texto o audio, manteniendo continuidad narrativa, consistencia semántica y adecuación al estilo comunicativo del emisor.

En el contexto de los deepfakes, estos modelos no generan por sí mismos contenido audiovisual falsificado. Sin embargo, desempeñan un papel habilitador crítico cuando se integran con tecnologías de síntesis de voz, sincronización audiovisual o generación multimodal. El contenido lingüístico que

producen actúa como base para construir mensajes hablados o audiovisuales atribuidos falsamente a una persona concreta.

Estos modelos permiten adaptar el discurso generado al tono, registro, estructura y estilo de comunicación de un emisor específico (institucional, profesional o técnico), lo que incrementa de forma significativa la credibilidad del contenido cuando se combina con audio o vídeo sintético.

Capacidades técnicas relevantes:

- Generación de textos estructurados adaptados al estilo comunicativo de perfiles institucionales o profesionales.
- Producción de contenido verbal destinado a su transformación en audio sintético mediante sistemas de clonación de voz.
- Ajuste del tono, nivel de formalidad y carga semántica del mensaje para reforzar su apariencia legítima.
- Integración con sistemas audiovisuales que alinean contenido verbal, voz y gestualidad sintéticas.

Impacto sectorial:

- **Smart Cities:** estos modelos permiten la generación de discursos, comunicados o instrucciones operativas falsas atribuidas a autoridades municipales, responsables de emergencias o mandos técnicos. Integrados con voz o vídeo sintético, estos mensajes pueden activar protocolos, desviar recursos o alterar decisiones operativas.
- **Salud:** los modelos autorregresivos pueden generar mensajes clínicos, aclaraciones diagnósticas o comunicaciones asistenciales que imitan el estilo de profesionales sanitarios. Combinados con audio o vídeo sintético, pueden suplantar a médicos en teleconsultas, introducir información falsa en el flujo asistencial o influir en decisiones clínicas.

5.2. Tipologías de deepfakes según el contenido manipulado

La clasificación de los deepfakes según el tipo de contenido afectado permite valorar su impacto específico y definir medidas de detección y prevención adaptadas a cada formato. Las tipologías más relevantes son las que afectan a imagen, audio, vídeo y contenido multimodal, dado que representan los principales canales empleados en procesos urbanos y sanitarios.

5.2.1. Deepfakes de imagen

Los deepfakes de imagen consisten en fotografías o documentos gráficos creados y/o alterados mediante técnicas de inteligencia artificial, o generados íntegramente de manera sintética. Pueden incluir modificaciones mínimas destinadas a engañar a sistemas automáticos o manipulaciones extensas que afectan a la estructura completa de la imagen.

En Smart Cities, los deepfakes de imagen pueden alterar capturas de cámaras de tráfico, modificar señales de circulación, suplantar identidades en documentos gráficos o alterar evidencias empleadas en

sanciones administrativas. En Salud, pueden manipular imágenes clínicas utilizadas para diagnósticos, modificar informes visuales o generar documentación fraudulenta con apariencia legítima.

5.2.2. Deepfakes de audio

Los deepfakes de audio consisten en la reproducción sintética de la voz de una persona o en la manipulación de grabaciones reales para modificar su contenido. Debido a la confianza que suele depositarse en la comunicación verbal, su impacto es especialmente significativo.

En Smart Cities, audios manipulados pueden generar instrucciones falsas, alterar la coordinación de servicios críticos o inducir actuaciones incorrectas en servicios de emergencia. En Salud, la voz falsificada de un profesional puede introducir cambios en pautas clínicas, engañar a pacientes o comprometer la autenticidad de instrucciones asistenciales.

5.2.3. Deepfakes de vídeo

Los deepfakes de vídeo combinan manipulación visual, gestual y acústica en una misma secuencia audiovisual, dando lugar a un vídeo sintético que aparenta registrar una escena, una actuación o una comunicación real. Su credibilidad es elevada, lo que los convierte en una de las amenazas más complejas cuando se emplean en contextos probatorios o asistenciales.

En Smart Cities, pueden alterar grabaciones de vigilancia urbana, simular incidentes inexistentes o modificar actuaciones policiales registradas. En el sector Salud, pueden suplantar a profesionales en consultas telemáticas, modificar grabaciones clínicas o generar secuencias que aparentan registrar intervenciones médicas reales.

5.2.4. Deepfakes multimodales

Los deepfakes multimodales integran manipulación simultánea en imagen, vídeo, audio y texto, produciendo contenido completo y altamente persuasivo. Su capacidad para simular comunicaciones institucionales de forma íntegra los convierte en un vector de riesgo estratégico.

En los entornos urbano y sanitario, pueden recrear mensajes oficiales, simular ruedas de prensa, generar alertas falsas o alterar comunicaciones clínicas completas que incluyan voz, imagen facial, texto superpuesto y gestos coordinados.

6. Riesgos y escenarios de deepfakes en Salud y Smart Cities

La incorporación de deepfakes en los entornos urbanos y sanitarios introduce riesgos que trascienden la dimensión puramente tecnológica y se manifiestan como amenazas funcionales, operativas, jurídicas y reputacionales. En sectores donde la información audiovisual sustenta decisiones asistenciales, administrativas, de seguridad pública o de gestión de emergencias, la presencia de contenido manipulado puede desencadenar consecuencias graves incluso cuando las infraestructuras tecnológicas no han sido comprometidas directamente.

Este capítulo describe los principales riesgos transversales asociados a los deepfakes y presenta escenarios específicos para los ámbitos de Smart Cities y Salud. Su finalidad es proporcionar una base sólida para la evaluación de riesgos y para la definición de controles y medidas de mitigación proporcionadas al impacto potencial.

6.1. Riesgos transversales y sistémicos

Los riesgos transversales afectan simultáneamente a organizaciones públicas, operadores tecnológicos, entidades sanitarias y ciudadanía. Representan amenazas estratégicas que pueden comprometer la confianza institucional, la estabilidad operativa y la integridad de los procesos de toma de decisiones.

- **Pérdida de integridad y autenticidad del contenido audiovisual**

Los deepfakes generan incertidumbre sobre la fiabilidad de las imágenes, vídeos y audios empleados para respaldar decisiones institucionales. Cuando no es posible distinguir entre contenido real y manipulado, la base probatoria de procedimientos administrativos, clínicos y operativos se ve comprometida. Esta pérdida de certeza afecta a la validez de sanciones, diagnósticos, informes técnicos, auditorías y comunicaciones oficiales.

- **Riesgo reputacional y desinformación coordinada**

Los contenidos sintéticos pueden emplearse para suplantar a autoridades, difundir declaraciones falsas, manipular la percepción pública o simular actuaciones institucionales inexistentes. Su rápida propagación en redes sociales y canales digitales puede amplificar crisis, generar alarma o deteriorar la confianza en organismos públicos y sanitarios.

- **Suplantación de identidad y fraude**

La clonación de voces y la falsificación de rostros permiten replicar la identidad de responsables municipales, profesionales sanitarios, técnicos de emergencias o personal administrativo. Esto habilita accesos indebidos, autorizaciones fraudulentas, manipulación de usuarios vulnerables y ataques de ingeniería social avanzados.

- **Alteración de procesos dependientes de evidencias audiovisuales**

Los deepfakes pueden distorsionar imágenes o grabaciones fundamentales para la gestión urbana o la asistencia sanitaria. La manipulación de pruebas visuales puede favorecer fraudes en trámites administrativos, influir en decisiones clínicas o alterar la documentación asociada a procedimientos regulatorios.

- **Aumento de la superficie de ataque**

Los deepfakes amplían la superficie de ataque al permitir generar impactos operativos, jurídicos y asistenciales sin necesidad de comprometer infraestructuras tecnológicas, redes o sistemas de información. A diferencia de los ciberataques tradicionales, que explotan vulnerabilidades técnicas, los deepfakes actúan directamente sobre la confianza en el contenido audiovisual y en su valor como evidencia.

Esto permite influir en decisiones clínicas, activar protocolos de emergencia, alterar procedimientos administrativos o generar desinformación institucional mediante la introducción de contenido audiovisual falsificado en canales legítimos, incluso cuando los sistemas tecnológicos subyacentes no han sido vulnerados. Como consecuencia, la superficie de ataque se extiende más allá de los sistemas informáticos y alcanza los procesos de toma de decisiones, la comunicación institucional y la validez probatoria del contenido.

- **Impacto en resiliencia y continuidad de negocio**

Mensajes sintéticos que simulan alertas de emergencia, instrucciones operativas o decisiones clínicas pueden activar protocolos de respuesta no justificados, generando interrupciones del servicio, descoordinación entre unidades y desgaste de recursos críticos. Este tipo de impacto afecta directamente a servicios esenciales como movilidad, emergencias, atención sanitaria o gestión de infraestructuras críticas.

6.2. Riesgos en entornos de Smart Cities

Las Smart Cities dependen de un ecosistema complejo de sensores, cámaras, redes de comunicación, plataformas de gestión urbana y sistemas automatizados de toma de decisiones. En este ecosistema altamente interconectado, los deepfakes pueden introducir distorsiones que afectan tanto a la operación diaria como a la seguridad pública.

- **Manipulación de sistemas de videovigilancia urbana**

Los deepfakes pueden alterar grabaciones de cámaras de tráfico, transporte público o seguridad ciudadana. Esto permite ocultar incidentes, simular comportamientos inexistentes o modificar pruebas empleadas en investigaciones y/o procedimientos sancionadores.

- **Interferencia en sistemas de gestión de movilidad**

Imágenes manipuladas de matrículas, vehículos inexistentes, señales alteradas o escenas urbanas sintéticas pueden afectar a sistemas de reconocimiento automático, algoritmos de optimización del tráfico, peajes inteligentes y controles de acceso.

- **Suplantación de autoridades o servicios municipales**

Deepfakes de voz o vídeo pueden imitar a responsables municipales, cuerpos de seguridad o personal de protección civil, emitiendo instrucciones operativas falsas, órdenes inexistentes o mensajes institucionales fraudulentos.

- **Impacto en la comunicación de emergencias**

La circulación de mensajes falsificados que simulan alertas meteorológicas, evacuaciones, riesgos químicos o incendios puede provocar confusión, saturación de líneas de emergencia y movilización indebida de recursos.

- **Fraude en servicios digitales municipales**

La clonación de voz o la generación de vídeos sintéticos puede emplearse para engañar a centros de atención telefónica, autorizar operaciones no permitidas, suplantar beneficiarios o manipular mecanismos de identificación biométrica.

- **Manipulación de procesos probatorios y administrativos**

Los deepfakes pueden afectar expedientes urbanísticos, documentación gráfica, informes técnicos y cualquier procedimiento administrativo basado en evidencia audiovisual, incrementando la probabilidad de fraude y litigios.

6.3. Riesgos en el sector Salud

El sector Salud depende de datos clínicos especialmente sensibles, imágenes diagnósticas, comunicaciones asistenciales y registros audiovisuales que influyen de manera directa en la seguridad del paciente. En este contexto, los deepfakes introducen riesgos que pueden comprometer decisiones diagnósticas, terapéuticas, organizativas y de investigación, afectando a la calidad asistencial y a la confianza en el sistema sanitario.

- **Alteración de imágenes médicas y diagnósticos asistidos por IA**

La manipulación sintética de radiografías, tomografías, resonancias, ecografías o imágenes dermatológicas puede añadir o eliminar hallazgos clínicos relevantes. Estas alteraciones pueden

- **Suplantación de profesionales sanitarios**

- **Manipulación de telemedicina y comunicaciones asistenciales**

- **Generación de documentación clínica sintética**

- **Riesgos en investigación biomédica**

- **Impacto en la percepción del paciente y en la confianza asistencial**

Página 25 de 50

7. Principios de seguridad para la gestión de deepfakes

La protección frente a deepfakes en los sectores de Smart Cities y Salud requiere aplicar principios de seguridad que integren gobernanza, prevención, verificación, respuesta y trazabilidad. Estos principios derivan de marcos europeos como el Reglamento de Inteligencia Artificial, la Directiva NIS2, el ENS, el RGPD y las normas internacionales ISO/IEC 42001, 23894 y 27001. Su finalidad es asegurar que las organizaciones puedan identificar, mitigar y gestionar los riesgos asociados a la manipulación audiovisual mediante técnicas de IA generativa, garantizando un funcionamiento seguro, confiable y ético de los servicios públicos y asistenciales.

Los principios definidos en este capítulo actúan como un marco transversal que integra tanto medidas técnicas como organizativas. En ellos se incluyen, de forma explícita o implícita, aspectos como la definición de políticas de uso responsable de la inteligencia artificial, la limitación de acceso a herramientas capaces de generar contenido sintético, el uso de marcas digitales y mecanismos de procedencia, la detección de manipulación audiovisual, así como la capacitación de empleados y la concienciación de la ciudadanía. El desarrollo operativo y técnico de estas medidas se concreta en los capítulos posteriores, especialmente en los apartados de detección (Capítulo 8) y de prevención y mitigación (Capítulo 9).

I. Principio de autenticidad y veracidad del contenido

Las organizaciones deben garantizar que las imágenes, vídeos y audios utilizados en procesos urbanos y sanitarios provienen de fuentes confiables y no han sido alterados. Esto implica establecer mecanismos sistemáticos de autenticación del origen, verificación de metadatos, trazabilidad del contenido y uso de técnicas de preservación de evidencias digitales.

Para ello, deben aplicarse controles sistemáticos de autenticación del origen, verificación y protección de metadatos, trazabilidad del contenido y preservación de evidencias digitales. En este marco se integran de manera natural el uso de marcas digitales, mecanismos de procedencia del contenido y tecnologías de detección de manipulación audiovisual, que actúan como soporte técnico para reforzar la confianza en la autenticidad del material.

En Smart Cities, este principio resulta crítico para cámaras de tráfico, sistemas de movilidad, sensores urbanos y plataformas de seguridad ciudadana. En el ámbito de la Salud, es esencial para validar imágenes diagnósticas, registros audiovisuales clínicos, contenidos de telemedicina y cualquier material que influya en decisiones asistenciales.

II. Principio de integridad y no manipulación

Toda organización debe asegurar que el contenido audiovisual crítico permanece íntegro desde el momento de su captura hasta su utilización operativa o asistencial. La integridad no se limita a evitar modificaciones evidentes, sino a garantizar que no se han producido alteraciones, reconstrucciones sintéticas o manipulaciones parciales que puedan distorsionar el significado del contenido.

Este principio exige controles de integridad técnica, sellado electrónico, cifrado de extremo a extremo, gestión fiable de metadatos y mecanismos de monitorización capaces de detectar alteraciones intencionadas o accidentales. En entornos urbanos y sanitarios, la integridad del contenido es un requisito indispensable para la validez de procedimientos sancionadores, diagnósticos clínicos, auditorías regulatorias y toma de decisiones críticas.

III. Principio de transparencia y trazabilidad

Las organizaciones deben ser capaces de reconstruir y demostrar, de forma clara y verificable, el origen, procesamiento, custodia y transformación del contenido audiovisual utilizado en procesos críticos. La trazabilidad permite sostener la confianza institucional incluso cuando existen dudas sobre la autenticidad del material.

Este principio se materializa mediante metadatos protegidos y verificables, registros de captura y modificación, sistemas de auditoría automatizada y mecanismos de identificación o etiquetado del contenido sintético cuando proceda. La trazabilidad resulta especialmente crítica en historiales clínicos, procesos diagnósticos y sistemas de vigilancia urbana, donde el contenido audiovisual puede tener consecuencias jurídicas o asistenciales directas.

IV. Principio de supervisión humana significativa

La utilización de contenido audiovisual en la toma de decisiones debe estar siempre sujeta a una supervisión humana real y proporcionada al nivel de riesgo. Ningún sistema automatizado debe operar de forma autónoma cuando el impacto potencial de un error pueda afectar a la seguridad pública, a la salud de las personas o a derechos fundamentales.

Este principio implica que operadores, técnicos municipales y profesionales sanitarios dispongan de procedimientos claros para revisar contenido, identificar anomalías, validar resultados y suspender decisiones automatizadas cuando exista sospecha de manipulación sintética. La supervisión humana solo es efectiva cuando va acompañada de formación específica y criterios operativos claros que permitan reconocer indicios de deepfakes y actuar conforme a los protocolos establecidos.

V. Principio de minimización y protección de datos personales

La captura, almacenamiento y procesamiento de contenido audiovisual debe limitarse a lo estrictamente necesario. Los deepfakes amplifican el riesgo asociado a la exposición de datos biométricos (rostro, voz, gestos), lo que exige reforzar las medidas de protección derivadas del RGPD y la LOPDGDD.

Este principio exige reducir la exposición innecesaria de datos audiovisuales, aplicar políticas de retención mínima, emplear técnicas de cifrado y anonimización cuando sea posible y proteger especialmente aquellas fuentes que podrían utilizarse para clonar identidades. Su aplicación resulta crítica tanto en sistemas de videovigilancia urbana como en el uso de imágenes diagnósticas y comunicaciones asistenciales.

VI. Principio de defensa en profundidad

La protección frente a deepfakes requiere un enfoque de defensa en profundidad basado en múltiples capas de control. Estas capas incluyen autenticación reforzada del contenido, detección automática de anomalías, validación manual especializada, monitorización continua de canales digitales y medidas de contención ante sospechas de manipulación.

Dentro de este principio se integra también la necesidad de limitar y controlar el acceso a herramientas capaces de generar contenido sintético. Las organizaciones deben regular su uso mediante políticas internas de seguridad e inteligencia artificial, evitando que tecnologías de generación avanzada puedan emplearse sin supervisión o fuera de los fines autorizados.

VII. Principio de resiliencia operativa y continuidad del servicio

Las organizaciones deben garantizar que la aparición de contenido manipulado no interrumpe la prestación de servicios esenciales ni compromete la seguridad pública o asistencial. Esto implica anticipar escenarios de deepfakes en los planes de continuidad, definir procedimientos de verificación rápida, establecer canales de comunicación seguros en situaciones críticas e implementar redundancias en las fuentes de información audiovisual.

En el sector Salud, la resiliencia es clave para mantener la continuidad asistencial y la fiabilidad diagnóstica. En Smart Cities, resulta determinante para la gestión de la movilidad, la seguridad ciudadana y la protección civil.

VIII. Principio de buena gobernanza y asignación de responsabilidades

La gestión de deepfakes exige una gobernanza clara y roles definidos. Este principio establece que las organizaciones deben asignar responsabilidades específicas en materia de integridad audiovisual, detección de contenido sintético, gestión de incidentes y supervisión técnica. Asimismo, debe formalizarse una política de uso responsable de la inteligencia artificial y del contenido sintético, que defina criterios de autorización, uso permitido, controles y rendición de cuentas, en coherencia con el marco regulatorio aplicable.

Los roles esenciales incluyen:

- **Alta dirección:** aprobación del marco estratégico y asignación de recursos.
- **CISO / Responsable de Seguridad:** diseño de controles, gestión de riesgos y supervisión de incidentes.
- **DPO / delegado de Protección de Datos:** supervisión del tratamiento de datos biométricos y cumplimiento RGPD/LOPDGDD.
- **Responsables de plataformas urbanas:** validación de contenido procedente de cámaras, sensores y sistemas de movilidad.
- **Responsables clínicos y de calidad asistencial:** verificación de imágenes diagnósticas, telemedicina y documentación audiovisual.
- **Comités de IA y responsable del sistema de IA:** supervisión del uso seguro y ético de sistemas generativos, conforme al Reglamento de IA.

- **AESIA y autoridades competentes:** en casos en los que la regulación exija supervisión externa o notificación.

La definición clara de roles garantiza una respuesta estructurada, coherente y verificable, evitando vacíos de responsabilidad ante incidentes de manipulación audiovisual. Se puede hacer referencia a la *Guía de modelo de gobierno de IA en entornos de Smart Cities y Salud* para asegurar la correcta implementación de gobernanza organizativa de la inteligencia artificial.

8. Metodologías de detección y verificación de autenticidad de deepfakes

La detección y verificación de deepfakes requiere un enfoque sistemático que combine criterios técnicos, procedimentales y contextuales con el fin de garantizar que el contenido audiovisual utilizado en entornos urbanos y sanitarios mantiene la autenticidad necesaria para servir de soporte fiable en decisiones administrativas, probatorias, asistenciales y operativas. A diferencia de otros vectores de riesgo tecnológico, los deepfakes actúan sobre la percepción humana y sobre la confianza en la evidencia audiovisual, por lo que su identificación exige metodologías más complejas que el simple análisis automatizado de archivos. Este capítulo define un modelo metodológico integral, adecuado para organizaciones que gestionan contenido crítico, y que permite evaluar la autenticidad del material audiovisual con rigor, trazabilidad y consistencia institucional.

La metodología se estructura en cinco fases consecutivas que abarcan desde la incorporación inicial del contenido hasta la elaboración de un dictamen final y su integración en los procesos organizativos. Estas fases pueden ejecutarse de forma secuencial o adaptarse a la criticidad del caso, pero en todos los escenarios deben garantizar la transparencia, la reproducibilidad y la coherencia con los marcos regulatorios aplicables.

8.1. Enfoque general y principios metodológicos

La metodología se articula sobre un enfoque escalonado que incrementa progresivamente la profundidad del análisis. Cada fase añade un nivel adicional de verificación, comenzando por observaciones perceptivas y estructurales y culminando en análisis avanzados capaces de identificar patrones residuales propios de las tecnologías generativas descritas en el Capítulo 5. Este enfoque no solo mejora la fiabilidad del proceso, sino que garantiza la proporcionalidad del esfuerzo técnico, evitando tanto la infraevaluación de material crítico como el sobredimensionamiento innecesario del análisis en contenidos de bajo impacto.

Cuatro principios fundamentales orientan la metodología:

Antes de aplicar las fases, debe verificarse que los sistemas de origen, custodia y análisis cumplen las medidas de autenticidad e integridad descritas en los Principios del Capítulo 7, especialmente en relación con trazabilidad, supervisión humana significativa y privacidad en el tratamiento de datos biométricos.

I. Proporcionalidad y adecuación al riesgo

El nivel de análisis debe ajustarse al impacto potencial que una manipulación del contenido podría generar, de acuerdo con los escenarios sectoriales descritos en el Capítulo 6. Mientras que una imagen de carácter divulgativo puede requerir únicamente un cribado inicial, una radiografía empleada en un diagnóstico, una grabación de telemedicina o el vídeo de un accidente de tráfico integrado en un procedimiento sancionador requieren un análisis mucho más profundo, incluyendo verificación avanzada y correlación multimodal. Este principio se alinea con la Supervisión Humana Significativa (Capítulo 7.4), que exige capacidad real de intervención humana en sistemas de alto riesgo.

II. Trazabilidad y registro sistemático

Cada acción realizada durante el análisis debe quedar documentada de manera verificable. Este requisito responde al Principio de Integridad y Trazabilidad expuesto en el Capítulo 7.3, así como a las exigencias de auditoría del ENS, NIS2 y MDR/IVDR. La trazabilidad incluye el registro del origen, la cadena de transmisión, las herramientas utilizadas, las observaciones detectadas, los resultados de cada fase y la versión final del contenido analizado. La ausencia de trazabilidad invalida la solidez del dictamen, incluso en caso de que el análisis sea técnicamente correcto.

III. Reproducibilidad y consistencia

Los procedimientos empleados deben permitir que distintos analistas lleguen a conclusiones coherentes si aplican la metodología de forma adecuada. Este principio evita subjetividades, garantiza la estabilidad del proceso y proporciona seguridad jurídica. Se relaciona directamente con los estándares de preservación de evidencias digitales recogidos en ISO/IEC 27037.

Este principio se aplica a todos los procedimientos de detección, verificación, trazabilidad, análisis forense y toma de decisiones operativas descritos en la presente guía.

IV. Integración sectorial y contextual

La detección de deepfakes debe integrarse en los procesos operativos, asistenciales y administrativos propios de cada organización. En Smart Cities, esto implica coordinar la verificación con sistemas de videovigilancia, plataformas de movilidad, centros de control operativo, servicios de emergencias y sensores urbanos. En Salud, con PACS, RIS, historia clínica electrónica, plataformas de telemedicina, flujos asistenciales y sistemas de gestión de imágenes diagnósticas. Ningún análisis puede considerarse completo si no se inserta correctamente en su contexto funcional.

8.2. Fase 1: Registro, Contextualización y Clasificación del Contenido Audiovisual

La fase inicial constituye el fundamento del proceso. Un registro deficiente puede invalidar incluso la verificación técnica más sólida, ya que la autenticidad no puede evaluarse sin conocer la procedencia, el contexto y el propósito del contenido.

a) Registro formal del contenido y su procedencia

Todo contenido audiovisual que pretenda utilizarse en procesos críticos debe registrarse en el momento mismo de su recepción o captura, incorporando:

- **Identificación del origen:** cámara urbana con identificador oficial, modalidad diagnóstica, plataforma de telemedicina, proveedor externo o ciudadano.

- **Metadatos fundamentales:** fecha, hora, localización, dispositivo de captura, operador responsable.
- **Canal de transmisión:** red institucional, mensajería profesional, repositorio clínico, plataformas externas.
- **Generación de hash inicial:** el archivo se somete al cálculo de un hash criptográfico (p. ej., SHA-256) que actúa como huella digital. Cualquier modificación posterior se detectará comparando este valor.

Este registro inicial constituye el punto de referencia para cualquier verificación posterior y forma parte de la cadena de custodia del contenido.

b) Análisis contextual previo

El archivo debe situarse dentro de su contexto operativo. Esto incluye identificar:

- La finalidad para la que se utilizará (sanción, diagnóstico, investigación, teleconsulta, verificación de suceso urbano, comunicación institucional, etc.);
- El proceso o procedimiento del que forma parte;
- La relación con otros datos previos (estudios clínicos, secuencias urbanas, registros de sensores, logs del sistema);
- El impacto potencial que tendría un error en su autenticidad.

Por ejemplo, un vídeo utilizado como prueba en una investigación policial, o una imagen diagnóstica que fundamenta una decisión clínica, requieren una interpretación contextual precisa antes de iniciar cualquier análisis técnico.

c) Clasificación de criticidad

La clasificación por niveles permite ajustar el esfuerzo metodológico posterior:

- **Nivel bajo:** contenido informativo, divulgativo o auxiliar que no influye de manera directa en decisiones críticas.
- **Nivel medio:** contenido que contribuye al proceso de toma de decisiones, pero que no constituye evidencia primaria (consultas telemáticas no urgentes, imagen clínica no determinante, material urbano complementario).
- **Nivel alto:** evidencia determinante en diagnósticos, procedimientos sancionadores, actuaciones policiales, emergencias, decisiones clínicas críticas, comunicaciones institucionales oficiales, o registros audiovisuales exigidos por marcos regulatorios como MDR/IVDR y/o el RIA.

La profundidad del análisis posterior depende de esta clasificación.

8.3. Fase 2: Cribado inicial y evaluación perceptiva

La segunda fase se basa en evaluación perceptiva humana y no sustituye ni pretende reemplazar los análisis técnicos y forenses descritos en las fases posteriores. Esta fase introduce un primer nivel de análisis directo sobre el contenido audiovisual, orientado a identificar indicios visibles o perceptibles de manipulación. Aunque no requiere conocimientos forenses avanzados, sí demanda formación específica para reconocer patrones anómalos característicos de los deepfakes descritos en el Capítulo 5. Su finalidad es determinar si el contenido presenta señales de alerta que justifiquen un análisis técnico más profundo en fases posteriores.

a) Evaluación visual preliminar

El análisis visual constituye el núcleo del cribado inicial y se basa en la observación detenida de elementos que, en condiciones normales, mantienen una coherencia física, óptica y anatómica.

El revisor debe examinar:

- **Coherencia lumínica:** orientación e intensidad de sombras, reflejos especulares, homogeneidad de la iluminación facial o del entorno, continuidad entre fotogramas. Las discrepancias pueden deberse a generación sintética o recomposición de segmentos de imagen.
- **Integridad morfológica:** contornos del rostro, orejas, pelo y manos; deformaciones localizadas; ausencia o irregularidad de microexpresiones; parpadeo artificial o con cadencia anómala.
- **Naturalidad del movimiento:** sincronía entre la gestualidad y la voz; deslizamientos faciales; rigidez en zonas concretas; falta de interacción física convincente con objetos o con el entorno.
- **Estructura global de la escena:** presencia de artefactos, duplicación de patrones, inconsistencias en la perspectiva o en la profundidad de campo, halos en los bordes y distorsiones durante cambios bruscos de pose.

En el ámbito sanitario, el cribado visual requiere consideración adicional de:

- La continuidad anatómica y lógica entre cortes (en estudios DICOM),
- La consistencia de las densidades tisulares,
- La relación fisiológica entre estructuras vecinas,
- La presencia de texturas excesivamente uniformes o artificialmente suavizadas.

En ambos sectores, cualquier inconsistencia perceptiva constituye un motivo suficiente para elevar el análisis a la siguiente fase.

b) Evaluación acústica preliminar

En contenidos sonoros o audiovisuales con voz humana, el cribado auditivo ayuda a detectar anomalías propias de la síntesis vocal:

- Ausencia de respiraciones, ruidos ambientales naturales o micromodulaciones propias del habla humana;
- Voz excesivamente uniforme, con timbre plano y sin variabilidad emocional;
- Transiciones entre fonemas demasiado suaves o, por el contrario, abruptas;
- Resonancias improbables en el espectro.

La comparación intuitiva entre la voz emitida y la gestualidad labial es especialmente relevante en vídeos donde se sospeche manipulación de labios, sincronización artificial o sustitución facial.

c) Evaluación contextual preliminar

El contenido debe ser evaluado respecto a su contexto verificable:

- Concordancia con horarios, meteorología, tráfico o actividad registrada por sensores urbanos;
- Coherencia con datos clínicos del paciente, estudios previos o anotaciones asistenciales;
- Consistencia con la secuencia previa o posterior de un vídeo;
- Alineación con protocolos institucionales, eventos programados o información oficial disponible.

Una discrepancia contextual significativa indica que la autenticidad del contenido no puede asumirse sin una verificación técnica más profunda.

8.4. Fase 3: Análisis técnico intermedio

Esta fase introduce técnicas objetivas que permiten examinar el archivo más allá de la percepción humana. A diferencia del cribado inicial, requiere competencias técnicas básicas y acceso a herramientas de análisis digital. Su propósito es identificar incoherencias estructurales, alteraciones del archivo y manipulación de metadatos.

a) Inspección detallada de metadatos

El análisis de metadatos constituye una de las herramientas más efectivas de detección temprana. Incluye:

- Revisión de etiquetas EXIF, XMP y metadatos específicos del formato: dispositivo de captura, firmware, fecha y hora de creación, parámetros de exposición, geolocalización, versiones de software y estructura interna del archivo.
- Detección de metadatos ausentes o incompatibles: especialmente relevante en imágenes clínicas DICOM, donde la ausencia de información obligatoria constituye un indicio crítico de manipulación o reconstrucción sintética.
- Verificación de sellos de tiempo cualificados conforme a eIDAS, que permiten acreditar jurídicamente el momento exacto de creación del archivo.

- Comparación entre metadatos declarados y el contexto de generación: divergencias entre fechas reales y declaradas, dispositivos imposibles, localizaciones incoherentes o metadatos generados tras modificaciones no autorizadas.

El análisis metadatos debe documentarse íntegramente, dado que constituye evidencia fundamental en auditorías técnicas y regulatorias.

b) Verificación de integridad y procedencia

Esta etapa evalúa la autenticidad estructural del archivo:

- **Comparación del hash inicial con el hash actual:** cualquier diferencia indica que el contenido ha sido modificado después de su recepción.
- **Verificación de firmas digitales del dispositivo de origen:** cámaras municipales certificadas, dispositivos diagnósticos o sistemas de telemedicina pueden generar firmas que autentican la captura.
- **Revisión de credenciales C2PA (Content Provenance and Authenticity):** cuando existan, permiten reconstruir la historia completa del archivo, incluyendo cada transformación sufrida.
- **Contraste con registros de sistemas institucionales:** en Salud, PACS, RIS o EHR; en Smart Cities, sistemas de videovigilancia, sensores, logs de movilidad y plataformas de control operativo.

Una discrepancia entre la procedencia declarada y la trazabilidad real obliga a activar la Fase 4 de detección y verificación.

c) Análisis estructural del archivo

Este análisis estructural del archivo permite identificar alteraciones invisibles al ojo humano:

- Pruebas ELA (Error Level Analysis) para localizar compresiones desiguales;
- Análisis del bitrate y de la estructura interna del vídeo, detectando variaciones abruptas o codificaciones incoherentes;
- Identificación de patrones duplicados o artefactos en regiones determinadas, propios de recomposición o sustitución;
- Detección de desalineaciones entre audio y vídeo, cambios anómalos en la frecuencia de muestreo o manipulación de secuencias.

Esta fase permite descartar una proporción significativa de falsificaciones sin necesidad de análisis forense especializado.

Nota: Las técnicas ELA resultan menos efectivas con imágenes generadas por modelos de difusión actuales, que producen patrones de compresión homogéneos. Deben considerarse un indicador preliminar, no conclusivo.

8.5. Fase 4: Verificación avanzada y análisis especializado

La cuarta fase está destinada a contenidos de alta criticidad o a situaciones donde las fases anteriores no permiten emitir un juicio fiable. Requiere personal especializado, herramientas avanzadas y, en algunos casos, laboratorios forenses acreditados. Es la fase más exhaustiva del proceso y su solidez determina la validez final del dictamen institucional.

a) Análisis forense audiovisual profundo

Se aplican técnicas diseñadas para detectar huellas residuales que los modelos generativos no pueden eliminar por completo:

- **Análisis PRNU (Photo-Response Non-Uniformity):** permite verificar si la imagen procede realmente del sensor físico declarado, comparando el “patrón de ruido” único que caracteriza cada dispositivo.
Nota: la eficacia del análisis PRNU disminuye en imágenes muy comprimidas o altamente procesadas, como ocurre en cámaras IP urbanas y algunos sistemas de gestión de vídeo. En el ámbito sanitario, muchas imágenes diagnósticas también son sometidas a procesos de compresión, normalización o reconstrucción propios de la modalidad clínica, lo que puede limitar igualmente la aplicabilidad del PRNU. Por ello, esta técnica debe emplearse con cautela y siempre como complemento de otros mecanismos de verificación, especialmente en imágenes médicas.
- **Análisis de frecuencias espaciales y temporales:** identifica artefactos generados por GAN, modelos de difusión o sistemas multimodales, especialmente en textura de piel, cabello, microdetalles y áreas sometidas a reconstrucción.
- **Mapas de calor de manipulación:** obtenidos mediante herramientas especializadas que destacan zonas alteradas o generadas sintéticamente.
- **Análisis espectral de audio:** detección de firmas de vocoder, reconstrucciones neuronales, ausencia de ruido basal o irregularidades imposibles en grabaciones físicas.
- **Detectores avanzados basados en IA:** utilizados únicamente como indicadores complementarios, nunca como determinantes, debido a la carrera de armamento entre modelos generativos y detectores.

En escenarios con implicaciones probatorias o regulatorias, el análisis forense avanzado deberá realizarse en laboratorios acreditados conforme a ISO/IEC 17025 o equivalentes, garantizando validez jurídica y técnica.

b) Comparación y correlación con fuentes independientes

Este procedimiento es esencial en entornos operativos donde existen múltiples sistemas capaces de registrar eventos simultáneamente.

En Smart Cities, se contrastan:

- Cámaras cercanas o adyacentes;
- Sensores de tráfico y aforos;
- Sensores acústicos, ambientales y meteorológicos;
- Registros de movilidad urbana;
- Datos de servicios de emergencias.

En Salud, se examina la coherencia con:

- Estudios clínicos previos del mismo paciente;
- Exploraciones complementarias;
- Informes de laboratorio;
- Biomarcadores;
- Secuencias anatómicas esperables según el tipo de estudio.

La ausencia de correlación con fuentes independientes constituye un indicio crítico de manipulación intencional.

c) Validación cruzada en sistemas de registro

La autenticidad del contenido se verifica revisando:

- La existencia de un log de captura coincidente;
- Los registros de transmisión y almacenamiento;
- La trazabilidad en plataformas clínicas (pacs, ris) o urbanas (vms, soc, sistemas de movilidad);
- Coherencia entre la hora del archivo y los registros del sistema de origen.

Una discrepancia sistemática entre el archivo y los registros institucionales exige clasificar el contenido en un nivel de riesgo máximo y restringir su uso.

8.6. Fase 5: Conclusión, documentación y decisión operativa

Esta fase integra los resultados de todo el proceso para generar un dictamen que determine si el contenido puede utilizarse en los procedimientos previstos. La forma y calidad de este dictamen serán determinantes para auditorías, decisiones judiciales, procedimientos sancionadores y actuaciones clínicas.

a) Elaboración del dictamen técnico

El dictamen debe incluir:

- Descripción del origen y contexto, incluyendo la clasificación de criticidad.

- Resultados específicos de cada fase, con explicación detallada de anomalías o ausencia de ellas.
- Herramientas, técnicas y fuentes consultadas, incluyendo sus limitaciones.
- Valoración final estructurada, utilizando la escala:
 - Nivel A – Autenticidad confirmada
 - Nivel B – Autenticidad probable
 - Nivel C – Autenticidad en duda
 - Nivel D – Contenido no fiable

Esta clasificación determina las decisiones operativas y la forma en que el contenido podrá o no integrarse en procesos asistenciales, administrativos o sancionadores.

b) Decisión operativa

Según la valoración final, el contenido podrá:

- Integrarse en la historia clínica del paciente;
- Utilizarse como evidencia en expedientes sancionadores o procedimientos administrativos;
- Emplearse en investigaciones policiales o urbanas;
- Descartarse por falta de fiabilidad;
- Someterse a verificaciones adicionales;
- Activar un protocolo formal de incidente de ia (capítulo 9).

La decisión operativa debe tomarse por los responsables designados en la gobernanza institucional, garantizando responsabilidad y trazabilidad.

c) Retroalimentación y mejora continua

Cada caso analizado alimenta el sistema institucional mediante:

- Actualización de procedimientos operativos;
- Mejora de protocolos de análisis y custodia;
- Fortalecimiento de capacidades técnicas;
- Actualización de la formación del personal;
- Identificación de amenazas generativas emergentes y adaptación a ellas.

8.7. Capacidades organizativas necesarias

La correcta implementación de esta metodología requiere capacidades organizativas adecuadas, tanto técnicas como de gestión.

I. Competencias y formación

La organización debe diferenciar tres niveles de capacitación:

- Formación básica para personal generalista en señales de manipulación perceptiva.
- Formación técnica intermedia en análisis de metadatos, integridad y estructura del archivo.
- Formación especializada forense para personal responsable de la Fase 4.

II. Infraestructura técnica esencial

Se requiere disponer de:

- Sistemas de autenticación en origen (firmas digitales, C2PA);
- Repositorios seguros para custodia de evidencias;
- Herramientas validadas para análisis forense;
- Plataformas de gestión de evidencias audiovisuales;
- Sistemas clínicos y urbanos con trazabilidad garantizada (PACS, RIS, VMS, SOC).

III. Gobernanza y roles institucionales

La organización debe definir responsabilidades claras:

- Responsables de análisis técnico, operativo y clínico;
- Responsables de seguridad (ciso), protección de datos (dpo) y cumplimiento regulatorio;
- Responsables de la toma de decisiones operativas;
- Punto de contacto único para casos sospechosos, habitualmente un soc o unidad equivalente.

Esta gobernanza garantiza coherencia, trazabilidad y capacidad de escalado. Se puede hacer referencia a la *Guía de modelo de gobierno de IA* que desglosa con detalle la implementación de gobernanza de la IA dentro de las organizaciones.

9. Medidas de prevención y mitigación de deepfakes

Las medidas de prevención y mitigación constituyen el complemento necesario a la metodología de detección expuesta en el Capítulo 8. Su objetivo es reducir la probabilidad de que contenido sintético alcance procesos críticos, limitar el impacto operativo de incidentes derivados de deepfakes y reforzar la resiliencia institucional en los sectores de Smart Cities y Salud. Estas medidas se estructuran en tres ámbitos principales: organizativos, técnicos sectoriales y comunicacionales. Su correcta implementación requiere coherencia con los marcos regulatorios aplicables (RIA, NIS2, ENS, MDR/IVDR, RGPD, LOPDGDD) y con los principios de gobernanza detallados en el Capítulo 7.

9.1. Medidas organizativas y de gobernanza

Las medidas organizativas establecen el marco estructural que permite aplicar las capacidades técnicas de manera proporcional y controlada. Constituyen la primera línea de defensa institucional.

a) Política de autenticidad e integridad del contenido audiovisual

Toda organización debe disponer de una política formal que regule el ciclo completo del contenido audiovisual crítico. Esta política debe especificar:

- Los requisitos de captura, transmisión, validación y custodia del material audiovisual;
- La clasificación del contenido según su riesgo, alineada con la tipología descrita en la fase 1 (capítulo 8.2);
- Los dispositivos, canales y plataformas autorizados para la recepción y procesado de contenido audiovisual;
- Las reglas de conservación, retención, destrucción y auditoría asociadas al material audiovisual, en coherencia con el ens, nis2 y la normativa sanitaria.

La ausencia de una política clara incrementa la probabilidad de que contenido sintético ingrese inadvertidamente en procesos sensibles.

b) Definición de roles y responsabilidades

La gobernanza frente a deepfakes exige roles definidos y procesos claros de coordinación interna. Las responsabilidades de cada rol definidas en la *Guía de modelo de gobierno de IA* deben integrar:

- Funciones estratégicas (dirección, responsables clínicos, responsables de plataformas urbanas);
- Funciones operativas (equipos técnicos, operadores de centros de control, personal clínico asistencial);
- Funciones de seguridad (ciso, soc, responsables de ia);
- Funciones regulatorias y de privacidad (dpo, unidades de cumplimiento);
- Funciones especializadas (equipos forenses institucionales, especialistas en imagen médica o análisis audiovisual).

Esta estructura operativa garantiza que las decisiones relacionadas con autenticidad audiovisual se tomen por perfiles adecuados y conforme al marco normativo aplicable.

c) Formación, capacitación y sensibilización

La formación debe adaptarse al nivel de responsabilidad de cada colectivo:

- Formación generalista para personal administrativo, asistencial y operativo en señales básicas de manipulación audiovisual;
- Formación técnica intermedia para operadores, técnicos y personal de sistemas en análisis estructural, verificación de integridad y gestión de metadatos;
- Formación especializada para analistas forenses responsables de la fase 4 (capítulo 8.5), con capacidad para aplicar técnicas avanzadas de detección multimodal.

La sensibilización periódica reduce la probabilidad de que contenido sintético se integre en flujos críticos sin una verificación adecuada.

d) Protocolos de respuesta a incidentes por deepfakes

La organización debe disponer de procedimientos específicos integrados en su Plan de Respuesta a Incidentes, que incluyan:

- Criterios para clasificar la gravedad del incidente;
- Medidas inmediatas de contención y aislamiento del contenido afectado;
- Procedimientos de verificación urgente mediante la metodología del capítulo 8;
- Mecanismos de escalado interno (soc, dirección, responsables clínicos o de movilidad);
- Criterios de notificación a autoridades competentes (incibe, aepd, aesia) cuando exista obligación legal.

La capacidad de activar una respuesta coordinada reduce el impacto operacional, clínico y reputacional de los deepfakes. Además, los incidentes confirmados o probables deberán activarse como “incidente de IA” conforme al Reglamento de Inteligencia Artificial cuando impliquen afectación a sistemas de alto riesgo en salud, seguridad pública o gestión de infraestructuras críticas, siguiendo los requisitos de documentación, notificación y trazabilidad definidos por el RIA.

9.2. Medidas técnicas en entornos de Smart Cities

Las Smart Cities integran redes complejas de sensores, cámaras y plataformas digitales que requieren controles de autenticidad de extremo a extremo. La prevención técnica debe concentrarse en tres áreas: origen, transmisión y validación.

a) Autenticación en origen de dispositivos urbanos

Los sistemas de videovigilancia, cámaras de tráfico y sensores visuales deben incorporar:

- Firmas digitales embebidas en el propio dispositivo;
- Sellado de tiempo confiable conforme al marco eidas;

- Inclusión automática de metadatos protegidos que documenten ubicación, identificador del dispositivo, estado operativo y parámetros de captura.

La autenticación en origen es la defensa más eficaz frente a la introducción de contenido sintético durante la captura.

b) Protección de la transmisión y custodia

La transmisión de flujos audiovisuales urbanos debe realizarse mediante:

- Canales cifrados y autenticados (VPN, TLS reforzado);
- Verificación de integridad en tránsito;
- Restricciones de acceso basadas en principios de mínimo privilegio.

Cualquier alteración durante la transmisión debe ser detectable mediante comparación de hash o validación de firma.

c) Validación automatizada en plataformas de gestión urbana

Los sistemas de gestión urbana (plataformas de movilidad, VMS, centros de control) deben implementar:

- Validación automática de firmas y metadatos;
- Análisis de coherencia espacial y temporal entre cámaras adyacentes;
- Integración de detectores de anomalías que alerten de patrones visuales improbables.

Estas medidas reducen el riesgo de que contenido manipulado influya en decisiones operativas.

9.3. Medidas técnicas en el sector Salud

La prevención en entornos sanitarios debe priorizar la integridad de imágenes diagnósticas, comunicaciones asistenciales y registros audiovisuales integrados en historias clínicas.

a) Protección en origen y circuito diagnóstico

Las modalidades diagnósticas (TAC, RM, ecografía) deben:

- Firmar digitalmente cada estudio DICOM en el momento de captura;
- Generar metadatos completos (UID de estudio, UID del dispositivo, parámetros técnicos);
- Impedir la exportación no autorizada del contenido mediante controles de acceso físico y lógico.

El PACS debe verificar firmas y conservar íntegramente la trazabilidad.

b) Garantía de integridad en telemedicina

Las plataformas de teleconsulta y telemonitorización deben asegurar:

- Autenticación multidimensional del profesional y del paciente;
- Cifrado integral de vídeo, audio y chat;
- Registro de capturas o grabaciones con metadatos verificables;
- Almacenamiento seguro en repositorios clínicos oficiales.

Cualquier contenido externo incorporado debe marcarse como “origen no verificado” hasta su validación.

c) Análisis automático de calidad y consistencia

Los sistemas sanitarios deben incorporar mecanismos de:

- Verificación de metadatos DICOM obligatorios;
- Análisis estadístico de texturas e intensidades;
- Revisión de firmas y trazabilidad.

Estos controles permiten detectar alteraciones sintéticas antes de influir en diagnósticos.

9.4. Mitigación en comunicación institucional

Las comunicaciones institucionales son especialmente vulnerables a deepfakes dirigidos contra la reputación o la coordinación operativa.

a) Garantía de autenticidad en comunicaciones oficiales

Para garantizar la autenticidad en las comunicaciones oficiales, las organizaciones deben:

- Difundir mensajes únicamente a través de canales oficiales certificados;
- Utilizar marcas de agua visibles e invisibles y sellos digitales;
- Publicar versiones verificables de discursos, notas de prensa y alertas.

b) Monitorización y verificación de contenido externo

Debe mantenerse una capacidad de vigilancia activa en redes sociales y medios digitales para identificar contenido que suplantan autoridades o imitaciones de alertas públicas.

c) Procedimientos de respuesta

Ante una falsificación:

- Se debe emitir una comunicación oficial verificable;
- Aportar evidencias técnicas que demuestren la falsedad;
- Activar canales de comunicación confiable con la ciudadanía y medios de comunicación.

10. Recomendaciones para usuarios y profesionales

La interacción cotidiana con contenido audiovisual convierte a la ciudadanía, al personal técnico municipal y a los profesionales sanitarios en actores clave dentro del ecosistema de protección frente a deepfakes. Este capítulo establece recomendaciones concretas adaptadas a cada perfil, orientadas a reforzar la capacidad de detección temprana, validar fuentes, activar canales oficiales y evitar que la manipulación sintética comprometa la seguridad urbana, la continuidad operativa o la seguridad del paciente. Las orientaciones que se presentan se integran en los principios de veracidad, supervisión humana, trazabilidad y resiliencia operativa descritos en los capítulos anteriores.

10.1. Ciudadanía y usuarios de servicios públicos

La ciudadanía es un objetivo frecuente de campañas de manipulación audiovisual y desempeña un papel central en interrumpir la cadena de propagación de contenidos sintéticos. La comprensión básica de señales de manipulación y la adopción de hábitos de verificación constituye un componente esencial de la resiliencia social frente a la manipulación audiovisual por lo que es necesario reforzar los siguientes elementos:

- **Reconocimiento de señales básicas de manipulación**

La ciudadanía debe prestar atención a desajustes evidentes entre imagen y sonido, expresiones faciales incoherentes, ausencia de elementos ambientales naturales y comportamientos que no resulten plausibles en el contexto en que se presentan. Las variaciones bruscas en iluminación, sombras o sincronía labial constituyen indicios comunes de contenido sintetizado mediante IA.

- **Hábitos de verificación antes de compartir contenido audiovisual**

Cualquier material que genere alarma, que invoque autoridad institucional o que incluya instrucciones operativas debe contrastarse directamente a través de sedes electrónicas, comunicaciones oficiales verificadas o plataformas institucionales certificadas. La verificación independiente es especialmente necesaria cuando el contenido circula por redes sociales, mensajería privada o canales informales.

- **Uso prioritario de fuentes oficiales en situaciones críticas**

Información sobre emergencias, tráfico, salud pública, protección civil o actuación municipal debe comprobarse exclusivamente mediante los canales digitales de la administración, que incorporan mecanismos de autenticidad verificable. La consulta inmediata en estas fuentes reduce la probabilidad de daño asociado a la desinformación audiovisual.

10.2. Personal técnico municipal y operadores de Smart Cities

El personal técnico de Smart Cities trabaja en entornos donde la información audiovisual influye directamente en decisiones operativas y sancionadoras. Su rol exige aplicar verificaciones sistemáticas sobre todo contenido que se incorpore a los sistemas urbanos.

- **Verificación estructural y contextual del contenido recibido**

El material audiovisual procedente de cámaras urbanas, sensores visuales o aportaciones ciudadanas debe revisarse contrastándolo con otras fuentes disponibles, como cámaras adyacentes, registros de tráfico, datos de movilidad y condiciones meteorológicas. La ausencia de correlación entre estas fuentes exige activar los procedimientos de verificación establecidos en la metodología.

- **Validación de instrucciones operativas y comunicaciones internas**

Las comunicaciones por voz o vídeo recibidas durante operaciones urbanas deben tratarse con especial precaución. Ante cualquier instrucción que se aparte de los protocolos habituales, la identidad del emisor debe verificarse mediante canales secundarios autorizados, evitando la ejecución de órdenes basadas en material sintético.

- **Gestión segura del contenido que ingresa en sistemas municipales**

Todo archivo audiovisual que provenga de fuentes externas o no autenticadas debe marcarse como no verificado y someterse al proceso metodológico descrito en el Capítulo 8. La documentación sistemática de sospechas y su escalado a unidades competentes garantiza trazabilidad y permite activar mecanismos de preservación y análisis.

10.3. Personal sanitario

Los profesionales sanitarios dependen de imágenes diagnósticas, comunicaciones asistenciales y documentos audiovisuales especialmente sensibles a la manipulación. Su labor exige integrar criterios técnicos y clínicos en la detección de anomalías, incluyendo:

- **Identificación de inconsistencias en imágenes clínicas**

Cuando un hallazgo radiológico no se corresponde con la evolución del paciente, cuando aparecen texturas anormalmente uniformes o cuando faltan características fisiológicas propias de la modalidad, el archivo debe considerarse potencialmente manipulado. En estas situaciones, resulta obligatorio aplicar los procedimientos de verificación técnica establecidos en la metodología.

- **Validación de identidad en telemedicina y comunicaciones asistenciales**

Las interacciones por vídeo o voz con pacientes o profesionales deben iniciar con la confirmación de identidad mediante datos clínicos verificables o protocolos establecidos por el centro. Cualquier comunicación clínica que resulte inusual, contradictoria o descontextualizada requiere confirmación mediante un canal alternativo.

- **Uso prudente de la imagen y voz profesional en entornos públicos**

La exposición pública de vídeos, fotografías o grabaciones de voz de profesionales sanitarios incrementa el riesgo de suplantación mediante modelos generativos. Se recomienda limitar esta exposición y notificar los responsables de seguridad cualquier intento de uso indebido o suplantación detectado en canales públicos o privados.

- **Reporte inmediato de incidentes o sospechas**

Ante cualquier documento audiovisual que genere dudas sobre su integridad o autenticidad, el profesional debe activar los canales internos de notificación, garantizando que la posible manipulación no compromete decisiones diagnósticas, terapéuticas o de trazabilidad clínica. El reporte temprano constituye un elemento central para la resiliencia del sistema sanitario.

11. Referencias

- Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial (AI Act).
<https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A32024R1689>
- Reglamento (UE) 2016/679 (RGPD), de 27 de abril de 2016, relativo a la protección de datos personales.
<https://eur-lex.europa.eu/eli/reg/2016/679/oj>
- Directiva (UE) 2022/2555 (NIS2), relativa al nivel elevado común de ciberseguridad en la Unión.
<https://eur-lex.europa.eu/eli/dir/2022/2555/oj>
- Reglamento (UE) 2022/868 (Data Governance Act) sobre la gobernanza europea de datos.
<https://eur-lex.europa.eu/eli/reg/2022/868/oj>
- Reglamento (UE) 2022/2065 (Digital Services Act – DSA) sobre servicios digitales.
<https://eur-lex.europa.eu/eli/reg/2022/2065/oj>
- Ley Orgánica 3/2018 (LOPDGDD), de Protección de Datos Personales y garantía de los derechos digitales.
<https://www.boe.es/buscar/act.php?id=BOE-A-2018-16673>
- Esquema Nacional de Seguridad (ENS) – Real Decreto 311/2022, de 3 de mayo.
<https://www.boe.es/eli/es/rd/2022/05/03/311>
- Carta de Derechos Digitales (Gobierno de España, 2021).
https://www.lamoncloa.gob.es/presidente/actividades/Documents/2021/Carta_Derechos_Digitales.pdf
- NIST AI Risk Management Framework 1.0 (2023).
<https://www.nist.gov/itl/ai-risk-management-framework>
- ISMS Forum – Grupo de Inteligencia Artificial (GIA). (2025). Deepfakes: riesgos, casos reales y desafíos en la era de la IA.
<https://www.ismsforum.es/ficheros/descargas/deepfake-final1742458135.pdf>
- ISO/IEC 42001:2023, Artificial Intelligence — Management system.
<https://www.iso.org/standard/81230.html>
- ISO/IEC 23894:2023, Artificial intelligence — Risk management.
<https://www.iso.org/standard/77304.html>

- ISO/IEC 38507:2022, Governance implications of the use of artificial intelligence.
<https://www.iso.org/standard/56641.html>
- ISO/IEC 27001:2022, Information security management systems.
<https://www.iso.org/standard/82875.html>
- ISO 31000:2018, Risk management — Guidelines.
<https://www.iso.org/standard/65694.html>
- ISO 37120, ISO 37122 e ISO 37123, Sustainable cities and smart cities — Indicators.
<https://www.iso.org/committee/656906.html>

12. Otras referencias de interés

El presente apartado recoge un resumen de los principales temas abordados en la guía, así como la relación de productos y servicios mencionados en ella. Su finalidad es ofrecer una visión global de los ámbitos tratados y facilitar la identificación de posibles referencias o elementos relevantes para futuras consultas o ampliaciones documentales.

12.1. Lista de materias tratadas en la guía

- Uso de sistemas de inteligencia artificial, incluida la IA generativa, en entornos de Smart Cities y en el sector Salud.
- Riesgos asociados a la generación y manipulación sintética de contenido audiovisual (deepfakes) en servicios públicos y sanitarios.
- Marcos normativos europeos y nacionales aplicables a inteligencia artificial, protección de datos personales, ciberseguridad y resiliencia digital.
- Autenticidad, integridad y trazabilidad del contenido audiovisual utilizado en procesos urbanos, administrativos y asistenciales.
- Gobernanza de la inteligencia artificial y supervisión humana significativa en sistemas de alto riesgo.
- Impacto de los deepfakes en la seguridad ciudadana, la gestión urbana, la atención sanitaria y la confianza institucional.
- Metodologías de detección, verificación y análisis forense de contenido audiovisual.
- Medidas organizativas, técnicas y operativas para la prevención y mitigación de incidentes relacionados con contenido sintético.
- Derechos de la ciudadanía y de los pacientes frente al uso de sistemas automatizados y contenido generado por IA.
- Comunicación institucional segura y gestión de la desinformación audiovisual.
- Guía de modelo de gobierno de IA en Smart Cities y Salud.

12.2. Lista de productos mencionados en la guía

No se mencionan productos específicos en la guía.

12.3. Lista de servicios mencionados en la guía

- Servicios municipales de movilidad y gestión del tráfico.
- Sistemas de videovigilancia urbana y seguridad ciudadana.
- Plataformas digitales de gestión urbana y centros de control operativo.
- Servicios sanitarios de diagnóstico por imagen y gestión clínica digital.
- Plataformas de telemedicina y atención sanitaria remota.
- Servicios de emergencias, protección civil y coordinación interinstitucional.

Detección y prevención de deepfakes generados por la inteligencia artificial en los entornos de Smart Cities y Salud

Noviembre de 2025

Versión 1.0



Junta de Andalucía