Deepfakes, desinformación y amenazas híbridas.

Guía práctica para proteger la reputación en la era de la IA generativa.

SEPTIEMBRE 2025

Foro-ia

Índice.

Foro-ia

Introducción.
Nuevas amenazas en la era digital.

En los últimos dos años la IA generativa ha irrumpido con fuerza en el mundo empresarial. Son evidentes las oportunidades que representa para aumentar la eficiencia y la productividad de numerosas áreas funcionales o explorar nuevos modelos de relación con el cliente. Sin embargo, esta revolución también tiene su reverso y, en malas manos, esta tecnología se convierte en una potentísima herramienta para manipular, desinformar y lanzar ataques mucho más sofisticados contra personas, organizaciones e incluso gobiernos.

Así lo confirma *The Global Risks Report 2025* de World Economic Forum. Este informe prevé que la manipulación y la desinformación ocuparán el primer puesto de su *ranking* de riesgos globales de aquí a dos años y advierte del papel que la IA jugará en todo ello.

La IA generativa reduce las barreras de entrada para la producción y distribución de contenido fraudulento y la suplantación de identidad, como en el caso de los *deepfakes*. Pero, sobre todo, permite crear materiales cada vez más difíciles de distinguir de la realidad.

Aun así, hablar de IA generativa y *deepfakes* de forma aislada no refleja la complejidad del nuevo entorno de amenazas híbridas que se abre ante nosotros. Por el contrario, este tipo de contenido engañoso es solo una pieza que, combinada con otros instrumentos ya clásicos —como los ciberataques, las identidades sintéticas o las redes de bots, el fraude y la extorsión—, condiciona decisiones clave, desestabiliza mercados y erosiona la ya frágil confianza y estabilidad del tejido social.

En el ámbito empresarial, en particular, estos ataques están provocando pérdidas económicas millonarias y daños reputacionales que pueden tardar años en revertirse. Hablamos, por ejemplo, de audios que suplantan la voz de directivos para autorizar transferencias bancarias. O de vídeos sintéticos con declaraciones que nunca realizaron para influir en los consumidores. En una encuesta global realizada por Deloitte en 2024, uno de cada cuatro ejecutivos afirmaba que sus organizaciones habían sufrido al menos un incidente con *deepfakes* en los 12 meses anteriores.

Introducción.
Nuevas amenazas en la era digital.

Este escenario no solo plantea retos operativos o de seguridad. Supone una disrupción directa en el terreno de la comunicación y la gestión reputacional. Las áreas encargadas de proteger la imagen pública de las organizaciones se enfrentan ahora a amenazas inéditas. Y en este nuevo contexto, reaccionar no es suficiente: es imprescindible anticiparse.

Desde Foro IA, queremos contribuir a esa preparación ofreciendo conocimiento, criterios y herramientas que ayuden a los profesionales de Marketing, Comunicación y Experiencia de Cliente (MCX) a afrontar estos nuevos riesgos con mayor solvencia. No existen modelos cerrados ni soluciones universales, pero sí una convicción clara: comprender el fenómeno es el primer paso para actuar.

Esperamos que esta guía te sirva como punto de partida.

Comprendiendo las amenazas híbridas.

1

1.1.
De las amenazas clásicas a las amenazas híbridas.

A lo largo de la historia, las amenazas a gobiernos, instituciones y empresas han adoptado muchas formas: desde la guerra convencional y el sabotaje económico, hasta el espionaje político, el fraude financiero o la manipulación informativa.

Sin embargo, en las últimas dos décadas, estas amenazas han dejado de operar de forma aislada. Han aprendido a coordinarse, a camuflarse y a aprovechar la interconexión digital, dando lugar a un nuevo escenario más complejo y difícil de atribuir: el de las **amenazas híbridas**.

Según el Centro Europeo de Excelencia para la Lucha contra las Amenazas Híbridas (Hybrid CoE), este fenómeno persigue debilitar objetivos estratégicos —como gobiernos, empresas o infraestructuras críticas— mediante **una mezcla de tácticas tecnológicas, informativas y psicológicas**. De este modo, se utilizan simultáneamente ciberataques, desinformación, presión política, chantaje económico, sabotaje o tecnología IA para erosionar la confianza, manipular la opinión pública o desestabilizar el funcionamiento de una sociedad.

Este último punto es especialmente relevante cuando hablamos de amenazas híbridas y nos sirve para recordar que un ciberataque o una noticia falsa que se producen de manera aislada no entrarían en esta categoría. Solo conforman una amenaza híbrida si se integran en una ofensiva más amplia y combinada contra un objetivo.

Por otro lado, la eficacia de este tipo de ataques reside precisamente en la ambigüedad, ya que logran permanecer bajo el umbral que desencadenaría una respuesta convencional, mientras provocan un impacto acumulativo.

Casos como el *deepfake* de voz que permitió robar 35 millones de dólares en 2021 o el que un par de años antes consiguió robar un cuarto de millón de dólares al CEO de una compañía energética evidencian que **la sofisticación de estas amenazas está en rápida evolución**. Y en este contexto, la irrupción de la IA generativa ha marcado un punto de inflexión.

«Las **amenazas híbridas** son actividades perjudiciales que se planifican y ejecutan con una **intención maliciosa**.

Su objetivo es **debilitar un objetivo**, como un Estado o una institución, a través de una combinación de distintos métodos. Estos incluyen la **manipulación de información, ciberataques, influencia o coerción económica, maniobras políticas encubiertas, diplomacia coercitiva o amenazas de fuerza militar**.

El concepto de amenazas híbridas abarca un amplio espectro de actividades dañinas con diversos propósitos, que van desde **operaciones de influencia e interferencia hasta la guerra híbrida**.»

Hybrid CoE
The European Centre of Excellence for Countering Hybrid Threats

1.2.
IA generativa, un salto cualitativo para las amenazas híbridas.

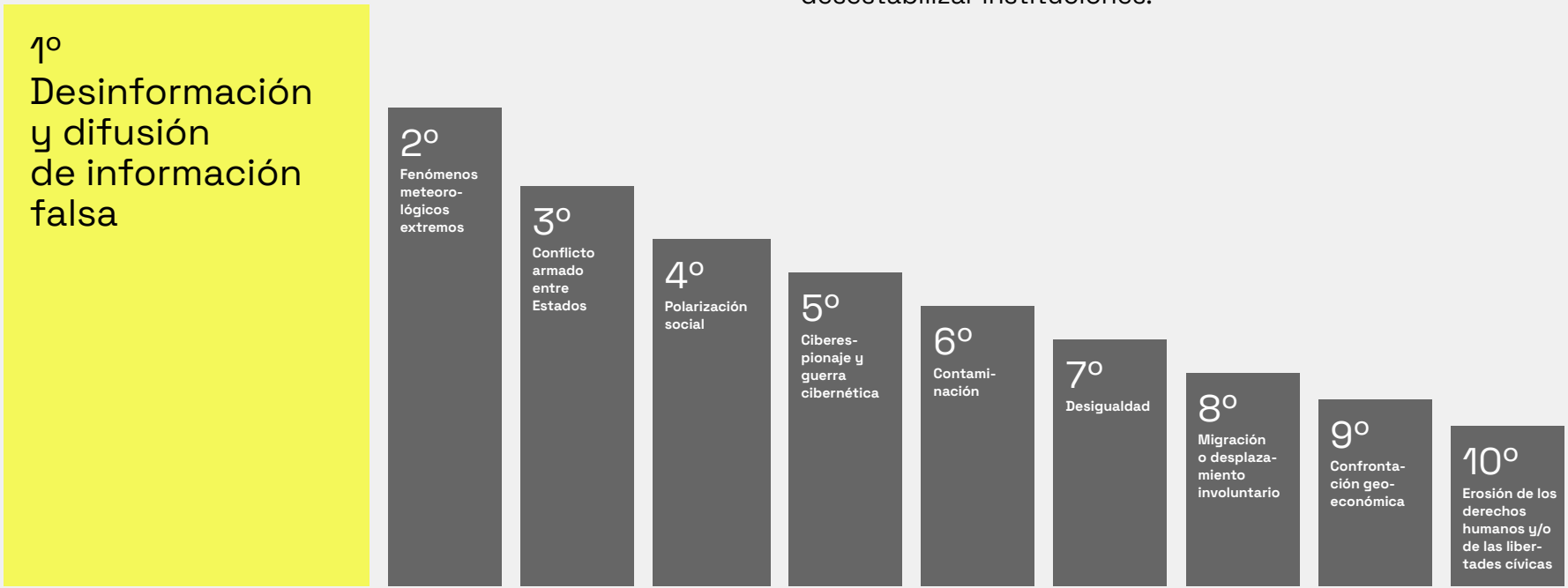
La IA generativa ha intensificado el fenómeno de las amenazas híbridas, haciéndolas más rápidas, más difíciles de detectar y mucho más escalables.

Mientras que hace apenas unos años los ataques a través de estrategias de desinformación requerían intervención humana intensiva —desde redactores hasta diseñadores de contenidos falsos— hoy basta con unos pocos comandos en un modelo de IA para generar *deepfakes* convincentes en cuestión de minutos (más sobre estos en el apartado 2.2), difundir narrativas falsas con bots automatizados o lanzar ciberataques de forma masiva con mínima inversión.

Además, la apertura y disponibilidad de modelos de IA de código abierto elimina las barreras de entrada, de modo que atacantes con poca experiencia técnica pueden acceder a herramientas avanzadas y replicar ataques que antes estaban limitados a servicios de inteligencia o cibercriminales de alto nivel. Esto está provocando una **explosión de incidentes** en sectores como el financiero, el sanitario o el de los medios de comunicación.

Principales riesgos percibidos (2024–2025).

El riesgo no es hipotético. Informes como el **Global Risks Report del Foro Económico Mundial**, señalan que la desinformación generada por IA es uno de los riesgos de mayor crecimiento a corto plazo, con capacidad real de alterar elecciones, influir en mercados financieros o desestabilizar instituciones.



Foro-ia

Fuente: World Economic Forum. *Global Risks Perception Survey 2024–2025.*

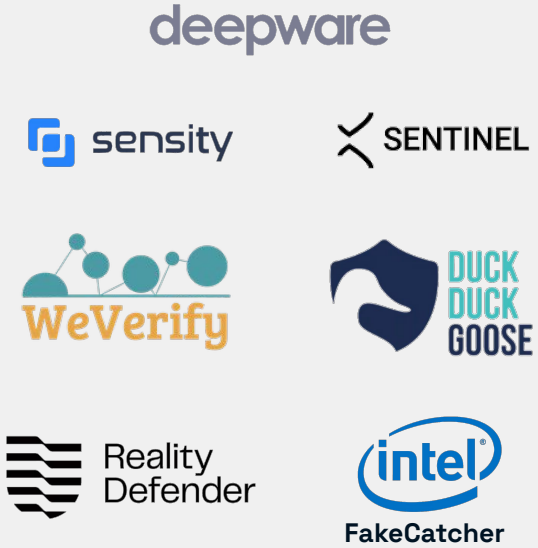
1.3.
IA generativa, también parte de la solución.

Venimos hablando de cómo la IA generativa amplifica el alcance y gravedad de las amenazas híbridas, pero no podemos terminar este capítulo sin recordar que esa misma tecnología puede convertirse en una aliada clave para contrarrestarlas.

De este modo, existen **herramientas capaces de detectar noticias falsas**, examinando el lenguaje utilizado, rastreando la fuente original, comprobando la coherencia con hechos conocidos o midiendo el grado de polarización emocional de un mensaje.

Por su parte, aquellas **especializadas en identificar *deepfakes*,** analizan aspectos técnicos como la sincronización labial, los movimientos faciales o los metadatos de los archivos. Mientras, los **sistemas de monitorización de redes sociales** buscan patrones sospechosos de viralización.

Por supuesto, la IA como herramienta para detectar y combatir las amenazas híbridas no es infalible —puede reproducir sesgos, arrojar falsos negativos o positivos—, pero sí que puede servir de filtro inicial sobre el que aplicar, después, el criterio y la supervisión humanos.



Algunas de las principales herramientas de IA para detección de *deepfakes*.

Amenazas híbridas en el entorno empresarial.

2

2.1.
Amenazas híbridas en el entorno corporativo.

En *The Landscape of Hybrid Threats: A conceptual model*, el Joint Research Centre of the European Commission (JRC) y el European Centre of Excellence for Countering Hybrid Threats (Hybrid CoE) desarrollan un modelo conceptual para caracterizar las amenazas híbridas y facilitar su gestión.

Definen con ese fin varios **pilares fundamentales**: el actor o atacante y sus objetivos estratégicos; el *target* u objetivo del ataque (ya sea un individuo, una empresa, una institución o un país); las herramientas utilizadas para influir, manipular, dañar o desestabilizar al objetivo; el dominio o ámbito en el que se produce el ataque; y, por último, las fases en las que se desarrolla la campaña híbrida.

Aunque este marco se diseñó inicialmente para apoyar a Estados y organismos públicos, nos proporciona un punto de partida para reflexionar y analizar los riesgos emergentes en el entorno corporativo y, en particular, las amenazas híbridas que se producen en el ámbito de la comunicación y la reputación corporativa.

Elementos de los ataques híbridos en el ámbito de la comunicación corporativa.

2.1.1.
Atacantes y motivaciones.

En el contexto corporativo, las amenazas híbridas tienen en muchos casos una base económica: **obtener ventajas competitivas, influir en mercados o debilitar a un competidor**.

Aún así, algunos ataques trascienden lo empresarial y adquieren tintes geopolíticos, especialmente cuando se dirigen contra compañías emblemáticas de un país o con alto valor simbólico. En estos casos, una empresa puede convertirse en objetivo más por lo que representa que por lo que produce.

Con estos objetivos, los ataques pueden proceder de:

Competidores directos.

Buscan socavar la posición de una empresa.

Hacktivistas y grupos afines.

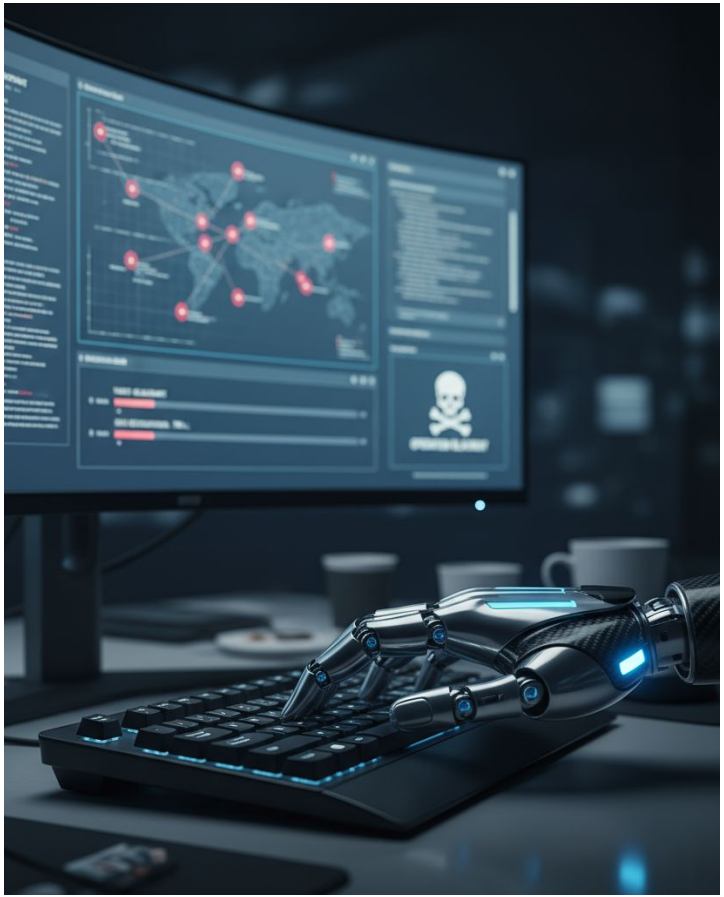
Operan de forma autónoma o esponsorizada, muchas veces con una causa o ideología como bandera.

Actores estatales.

Utilizan estos ataques como arma geopolítica.

Actores indirectos.

Medios de comunicación, asociaciones o incluso clientes. Acaban participando en la propagación del ataque.



2.1.2.
Target u objetivo del ataque.

En una amenaza híbrida, el objetivo no siempre es evidente ni único. De hecho, aunque estemos centrándonos en el contexto empresarial, el ataque no siempre se limita a la organización y puede extenderse a **personas relevantes en la organización y empleados en general, clientes, proveedores o aliados estratégicos**, amplificando así su impacto y presión.

Además, en una economía global y digitalizada, cualquier actor puede convertirse en puerta de entrada, amplificador o víctima colateral. Muchas veces, **una pyme puede ser utilizada como punto de entrada para dañar a un cliente mayor** o acceder a infraestructuras críticas. En otras ocasiones, la empresa puede no ser ni siquiera el objetivo final, sino un medio para enviar un mensaje al Estado que la abandera o a los mercados en los que opera. Es por ello que las grandes corporaciones están revisando y ampliando los sistemas de control y seguridad de todos sus *partners*.

Por otro lado, existen **sectores especialmente vulnerables** a este tipo de ataques:

Energía	**Finanzas**	**Transporte y logística**
Telecomunicaciones	**Tecnología**	**Sanidad**

CASO REAL — QANTAS AIRLINES

Cuando un pequeño proveedor es el punto débil de la cadena.

Este mismo verano, un ataque a Qantas Airlines expuso los datos personales de hasta 6 millones de clientes. Sin embargo, este no se dirigió directamente a la compañía, sino a uno de los *call centers* con los que trabajaba.

Aunque ni las autoridades ni la empresa han llegado a confirmar de manera oficial el método a través del que se produjo el ataque, todo apunta a que fue ejecutado a través de técnicas de ingeniería social, en concreto *vishing* (*phishing* por voz).

Unos días antes, las autoridades estadounidenses habían advertido que el sector aéreo estaba siendo objetivo de un grupo conocido como Scattered Spider, que suplantaba a empleados o contratistas para engañar a los departamentos de soporte técnico de TI y obtener acceso a sistemas de terceros.

2.1.3.
Herramientas y canales.

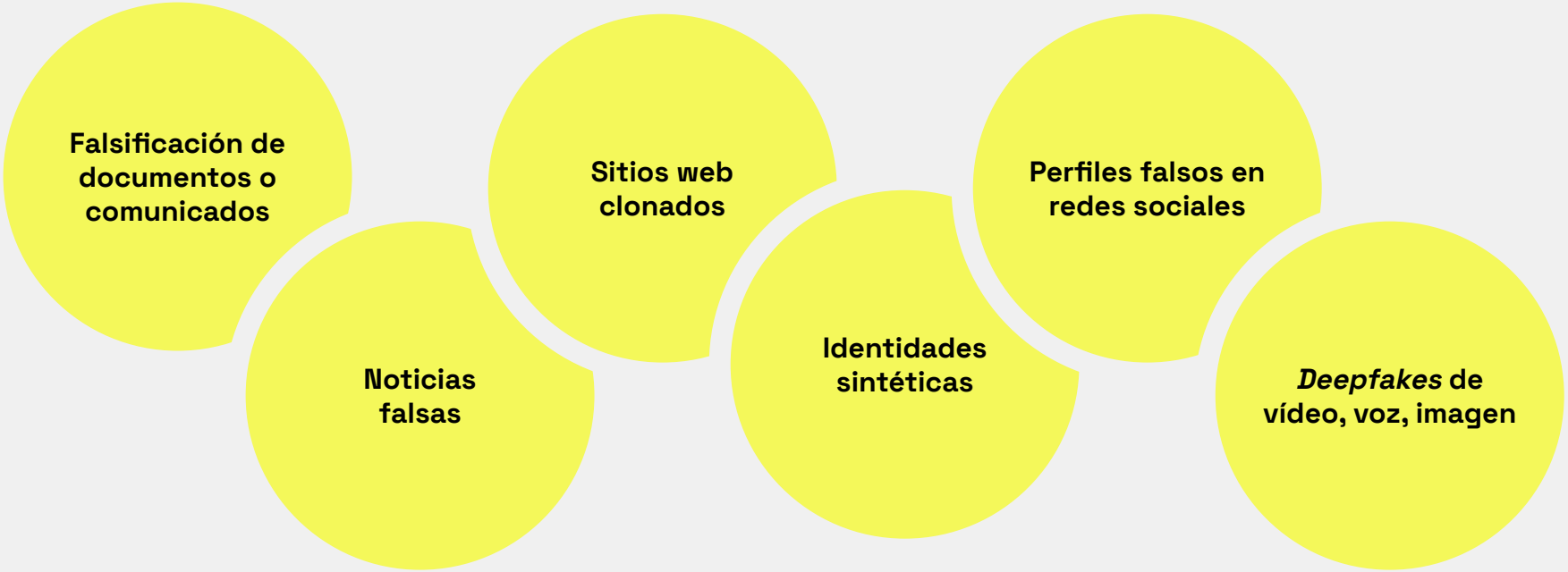
Las amenazas híbridas suelen dirigirse primero a los puntos más vulnerables o estratégicos de una organización. A partir de ahí, escalan de forma secuencial o simultánea, conforme se van alcanzando los objetivos marcados. Su capacidad de combinación —y mutación— las hace especialmente difíciles de anticipar y contener.

Se despliegan de forma coordinada en **múltiples dominios** —como el informativo, el económico o el cibernético— y combinan **herramientas físicas, digitales y psicológicas**.

Muchos de estos ataques se desencadenan en canales digitales públicos como las redes sociales. Sin embargo, otros tantos suceden en entornos no supervisados o de difícil monitorización, como aplicaciones de mensajería privadas o espacios de *dark social*. Esto hace que los marcos tradicionales de detección y respuesta resulten insuficientes y obliga a las empresas a adoptar **enfoques más integrados, multidominio y proactivos**.

2.1.3.
Herramientas y canales.

En estos ataques encontramos **una mezcla de recursos clásicos con nuevas capacidades tecnológicas.** Estas herramientas, combinadas con ciberataques clásicos e ingeniería social, fraudes comerciales, filtraciones o boicots, buscan desinformar, suplantar, extorsionar o simplemente sembrar el caos.



En una encuesta global realizada por Deloitte en 2024, **1 de cada 4 ejecutivos** aseguró haber sufrido al menos un ataque de *deepfake* en los 12 meses previos.

2.2.
Deepfakes.

Los *deepfakes* son una forma de manipulación avanzada que utiliza la IA para crear vídeos, audios, imágenes e incluso textos que son falsos pero que parecen auténticos y resultan convincentes. Emulan con gran precisión a personas reales —sus rasgos faciales, sus gestos, su voz— para hacernos creer que esa persona hizo o dijo algo que nunca sucedió, llevando al límite nuestra capacidad para distinguir el contenido real del manipulado.

El término fue acuñado en 2017 por un usuario de Reddit que publicó los primeros vídeos alterados de celebridades usando esta técnica. Es una combinación de «*fake*» (falso) y «*deep*», que hace referencia a «*deep learning*» (aprendizaje profundo), la tecnología que está detrás de este tipo de contenidos sintéticos.

En 2024, **una de cada dos empresas a nivel mundial informó de incidentes de fraude con *deepfakes***, según *The Deepfake Trends 2024* de Regula, y algunas proyecciones, como las que publica World Economic Forum, advierten de que para 2026 hasta el 90% del contenido *online* podría ser sintético o generado por IA.

El *deepfake* del Papa Francisco luciendo un abrigo estilo Balenciaga (2023) disparó el debate público acerca de los riesgos de la IA generativa.

Tipos de *deepfakes*.

Estos formatos pueden ser utilizados de manera independiente, pero también combinados, como por ejemplo en la creación de una página web fraudulenta.

Vídeo

Reemplaza el rostro de una persona por el de otra, simulando que esta habla o actúa en la escena. Con técnicas de *face swapping* y *reenactment* se pueden sincronizar expresiones faciales y movimientos labiales con gran realismo.

Imagen

Crea fotografías e imágenes estáticas realistas de personas que nunca existieron o personas reales en situaciones ficticias.

Voz

Clona la voz de una persona y genera un audio falso con el mismo tono y timbre. Se pueden fabricar grabaciones o realizar llamadas telefónicas donde alguien aparentemente dice algo que en realidad nunca ha dicho.

Audio

Genera textos sintéticos que imitan el estilo de redacción de una persona o institución. Se pueden producir emails, comunicados de prensa, mensajes en redes sociales u otros documentos.

2.3.
Principales tipos de amenazas híbridas y casos de uso.

2.3.1.
ESCENARIO 1
—
Suplantación de identidad para acceso a datos o pagos fraudulentos

2.3.2.
ESCENARIO 2
—
Deepfakes para manipulación de audiencias.

2.3.3.
ESCENARIO 3
—
Noticias falsas para manipular productos o mercados

2.3.1 ESCENARIO 1

Suplantación de identidad para acceso a datos o pagos fraudulentos.

Modalidad de fraude corporativo que replica de forma hiperrealista la apariencia y/o la voz de directivos, responsables financieros o figuras de autoridad dentro de una organización para obtener acceso a sistemas críticos e información sensible o lograr transferencias económicas fraudulentas.

Foro-ia

Este tipo de ataques utilizan ***deepfakes* de voz y vídeo**, combinados con credenciales o documentación falsas, para llegar a empleados con capacidad de decisión o acceso a sistemas financieros, como responsables de tesorería, administración o IT.

Son **dirigidos** (*spear phishing*) **y altamente personalizados** en función del perfil del receptor, y suelen producirse en **canales privados o semiprivados** que escapan al control de los sistemas tradicionales de ciberseguridad y monitorización: servicios de mensajería instantánea, correos personales, plataformas de videollamada o incluso llamadas directas con voces sintéticas. Por eso mismo, a los directivos se les suele recomendar evitar el uso de herramientas de mensajería instantánea generales y utilizar aquellas recomendadas y protegidas dentro de su entorno corporativo.

Para las compañías, este tipo de ataques pueden generar pérdidas económicas millonarias en cuestión de minutos. Aunque más allá del impacto financiero, este tipo de incidentes **erosiona la confianza interna, obliga a revisar los protocolos de seguridad y puede conllevar sanciones regulatorias** si se demuestra negligencia en la protección de datos o el control de accesos.

2.3.1 CASO REAL — ARUP

El CFO sintético que le costó 25 millones de dólares a Arup.

En mayo de 2024, la multinacional de diseño e ingeniería Arup reconoció haber sido víctima de un ataque que llevó a uno de sus empleados de Hong Kong a transferir más de 25 millones de dólares a unos estafadores.

El empleado no hizo más que seguir instrucciones tras una videollamada en la que creyó ver y oír juntos al director financiero de la compañía y otros miembros de su equipo, cuando en realidad eran solo avatares sintéticos.

Según la investigación policial posterior, el trabajador llegó a sospechar de un intento de *phishing* tras haber recibido un correo desde la oficina del Reino Unido de la empresa en el que se le indicaba la necesidad de realizar una transacción confidencial. Sin embargo, disipó sus dudas tras esa videollamada en la que sus interlocutores parecían totalmente reales.

Esta modalidad de *deepfake* «en tiempo real» amplifica aún más el potencial daño de este tipo de ataques, puesto que dificulta la verificación durante el evento mismo.

2.3.2 ESCENARIO 2

Deepfakes para manipulación de audiencias.

Creación de contenidos audiovisuales falsos en los que la marca o un alto cargo o portavoz de la compañía engaña al público para redirigirlo a una actividad fraudulenta. Puede ser un falso servicio por el que tengan que pagar una cantidad determinada de dinero o un formulario de suscripción o registro que permita robarle sus datos.

Foro-ia

Este tipo de contenidos falsos pueden presentarse como entrevistas, comunicados oficiales o testimonios falsos.

Aparecen en diferentes tipos de canales digitales, como redes sociales, según el caso, **como contenido orgánico o también como campañas publicitarias**.

Este último caso —promoción mediante campañas publicitarias— es especialmente difícil de detectar y monitorizar, debido a las condiciones de funcionamiento de estas campañas y las propias plataformas, y las limitaciones técnicas actuales.

Aparte del daño reputacional que pueden suponer para la marca y el propio directivo o persona suplantada, este tipo de ataques **entraña riesgos legales para las compañías** si estas no actúan o su respuesta se demora demasiado —por infracción de las normas de protección al consumidor o de protección de datos, posibles denuncias de los afectados o similares, según el país en que opere la empresa—.

2.3.2 CASO REAL — MARTA ORTEGA

El falso proyecto financiero de Marta Ortega.

En diciembre de 2023, RTVE desmintió haber emitido un vídeo que circulaba en redes sociales y en el que se veía a la presentadora Ana Blanco hablando de un supuesto proyecto financiero creado por Amancio Ortega y el Banco de España.

El montaje suplantaba la identidad de la presentadora, así como de la periodista Almudena Guerrero, el gobernador del Banco de España, Pablo Hernández de Cos, y la presidenta del grupo Inditex, Marta Ortega, para invitar a la ciudadanía a utilizar una plataforma con la que podrían obtener rendimientos de hasta 26.000 euros al mes con una inversión mínima.

No fue un hecho aislado. Meses después, en marzo y en septiembre de 2024, la cadena volvió a advertir de la difusión de otros dos vídeos de contenido similar que involucraban siempre a la presidenta de Inditex junto a los presentadores de RTVE Ángel Pons y la reportera del Telediario Clara Sedano, en un caso, y al presentador del Canal 24 Horas Lluís Guilera, en otro.

Tras analizar las imágenes y el audio del vídeo, VerificaRTVE concluyó que era falso. Para realizarlo, se había realizado un vídeo original de Marta Ortega correspondiente a la Junta General de Accionistas de Inditex celebrada en julio de 2022 en Arteixo (A Coruña). Por supuesto, en el original no hacía ninguna referencia al *software* financiero que le atribuían en la grabación de Facebook.

2.3.3 ESCENARIO 3

Noticias falsas para la manipulación de productos o mercados.

Fabricación o distorsión deliberada de información para crear una percepción falsa sobre un producto o servicio, la salud de una empresa o una tendencia del mercado. Busca inducir una reacción rápida por parte de consumidores o inversores, para desestabilizar al objetivo y, a menudo, producir ganancias financieras para los atacantes.

Este tipo de ataque contempla la difusión de noticias, rumores o análisis falsos o engañosos a través de diferentes canales, como redes sociales o sitios web de noticias falsas.

En ocasiones, estas técnicas **se combinan con actividades comerciales** específicas como la venta o compra masiva de acciones.

Es un tipo de amenaza dirigida a un amplio segmento de la población en lugar de a individuos específicos, con **capacidad de monitoreo parcial** dependiendo de las fuentes digitales en las que ocurre.

2.3.3 CASO REAL — EL PENTÁGONO

Una explosión que nunca ocurrió pero que agitó los mercados.

El lunes 22 de mayo, poco después de las 10:00 a. m. (hora local), una fotografía que mostraba una supuesta explosión frente al edificio del Pentágono comenzó a circular en redes sociales. En la imagen, una gran nube de humo negro se alzaba junto a lo que parecía ser una instalación gubernamental, acompañada de mensajes que sugerían un ataque.

Poco después de la difusión de la imagen, el Departamento de Defensa de EE. UU. confirmó que no se había producido ninguna explosión cerca del Pentágono. La imagen había sido generada con IA.

Sin embargo, antes de que las autoridades pudieran desmentir la noticia, los mercados ya habían reaccionado. El índice Dow Jones —que refleja el valor de grandes empresas como Coca-Cola o Apple— bajó unos 80 puntos y el índice S&P 500 registró una variación a la baja de un 17%. Todo ello en 3-4 minutos.

Aunque los mercados se recuperaron rápidamente, el episodio bastó para evidenciar lo extremadamente frágil que puede ser la estabilidad financiera ante este tipo de amenazas.

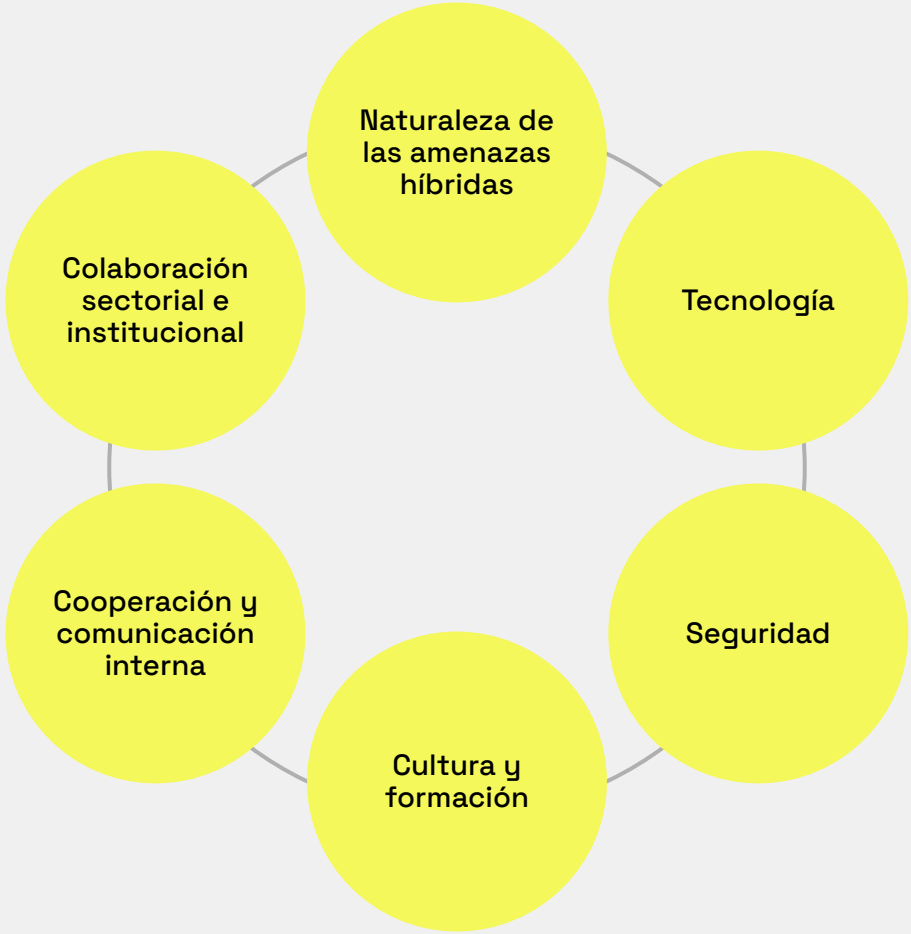
Retos organizativos y tecnológicos.

3

3.1.
Principales retos ante las amenazas híbridas.

Aunque el escenario de las amenazas híbridas y su complejidad creciente puedan resultar abrumadores, en ningún caso debemos quedarnos con la idea de que son inevitables o imposibles de combatir.

Las empresas no están indefensas frente a este tipo de riesgos. Al contrario, pueden y deben desarrollar capacidades organizativas y tecnológicas para enfrentarlas. Esto exige una transformación profunda de cómo se entiende y se gestiona la seguridad corporativa.



3.1.
Principales retos ante las amenazas híbridas.

Naturaleza de las amenazas híbridas.

Uno de los mayores desafíos de este nuevo escenario es que muchas de estas amenazas **operan bajo el radar de detección tradicional** y esto dificulta enormemente su identificación temprana.

Una campaña de *deepfakes* puede iniciarse en primer lugar en canales cerrados (el llamado *dark social*, como WhatsApp o Telegram), a través de cuentas falsas, bots coordinados o páginas clonadas, donde escapan a los sistemas de monitorización convencionales.

Por otro lado, las consecuencias no siempre son inmediatas ni fácilmente visibles. Un *deepfake* viral puede sembrar la duda en el consumidor incluso si es desmentido horas después, y una narrativa falsa sobre una empresa puede instalarse en la conversación pública aunque no tenga una base factual y generar una erosión lenta y progresiva de la confianza en esa compañía.

Limitaciones tecnológicas.

La propia naturaleza de las amenazas híbridas hace que las herramientas tecnológicas actuales sean insuficientes para detectarlas.

Por otro lado, incluso cuando aparecen esas soluciones, habrá que tener en cuenta la capacidad de cada compañía para acceder a ellas. Si pensamos en un país de pymes como España, comprenderemos que **muchas organizaciones no disponen ni de los recursos ni de la formación necesaria** para utilizar estas herramientas con eficacia.

3.1.
Principales retos ante las amenazas híbridas.

Seguridad.

Esto obliga a mantener un **enfoque realista y dinámico de la seguridad.**

La protección total (100%, 24/7) no existe, pero sí es posible es construir una arquitectura que permita anticipar riesgos, responder con agilidad y mitigar el impacto reputacional, operativo y relacional que pueden generar estos ataques.

Para ello, resulta clave **dimensionar correctamente el nivel de exposición a este tipo de amenazas**, plantear escenarios realistas —como campañas de desinformación, suplantación de identidad de directivos, uso de *deepfakes* para generar crisis reputacionales o manipulación de narrativas de marca— y diseñar protocolos de respuesta adaptados a cada uno de ellos.

Cultura corporativa y formación continua.

La respuesta a las amenazas híbridas requiere una transformación cultural.

La formación y concienciación deben ser continuas, tanto internamente **(empleados, portavoces, directivos)** como externamente **(clientes, proveedores, *stakeholders* clave)**. Como ocurre en el ámbito de la ciberseguridad, la primera línea de defensa ante una campaña de desinformación no es tecnológica, sino humana.

Formar para un uso apropiado de la tecnología, pero también para facilitar el reconocimiento de manipulaciones informativas debe ser parte del ADN de cualquier organización que aspire a mantener su reputación y la confianza de sus públicos.

3.1.
Principales retos ante las amenazas híbridas.

Comunicación y cooperación interna.

Ahora bien, frente a las amenazas clásicas, este esfuerzo no puede recaer exclusivamente en los departamentos de comunicación o de ciberseguridad.

Muy al contrario, exige una **respuesta interdepartamental** en la que colaboren áreas como legal, relaciones institucionales, tecnología, atención al cliente y, por supuesto, la alta dirección.

Una empresa resiliente ante amenazas híbridas es aquella que articula una **inteligencia colectiva**, en la que cada función aporta una visión complementaria y contribuye a una estrategia de protección global.

Colaboración sectorial y presión institucional.

Por último, debemos ser conscientes de que el reto es tan grande que ninguna empresa puede enfrentarlo sola.

Se requiere una colaboración estrecha entre **actores públicos y privados** para compartir buenas prácticas y diseñar estrategias conjuntas, mecanismos de protección y normativas para hacer frente a estas amenazas.

Aquí, además, hay un reto específico que tiene que ver con **el rol de las plataformas** a través de las que se distribuyen este tipo de contenidos fraudulentos. De todo ello, hablamos en más profundidad en el capítulo 5 de este *whitepaper*.

La respuesta a las amenazas híbridas requiere una **transformación cultural**.
La primera línea de defensa no es tecnológica, sino humana.

3.2.
Las amenazas híbridas en el contexto de la pyme.

Aunque todas las organizaciones son potenciales objetivos de este tipo de ataques, existen diferencias notables entre cómo estos afectan y son gestionados por grandes empresas y pymes.

En general, las **corporaciones grandes disponen de mayores recursos financieros y humanos** para monitorear y contrarrestar estas amenazas. Suelen contar con departamentos de comunicación robustos, equipos dedicados de gestión de crisis, asesores legales y de relaciones públicas, e incluso herramientas tecnológicas avanzadas (monitorización 24/7 de redes, servicios de inteligencia digital, etc.). Gracias a ello, pueden *capear el temporal* con más eficacia, detectando antes las narrativas dañinas, respondiendo públicamente con comunicados oficiales o ruedas de prensa, movilizando apoyos de terceros influyentes y desplegando contramedidas (como puede ser colaborar con plataformas para eliminar contenido falso o poner en marcha acciones legales).

3.2.
Las amenazas híbridas en el contexto de la pyme.

En contraste, las **pequeñas y medianas empresas suelen operar con estructuras y equipos más reducidos**, que, a menudo, desempeñan múltiples funciones. Esto puede hacer que tarden en detectar el ataque y reaccionar, lo que favorece su propagación, o que lleven a cabo acciones insuficientes.

Tal y como se indicaba en capítulos anteriores, una pyme puede considerarse un medio para perjudicar a un objetivo mayor, con lo que no se debe despreciar el impacto potencial de cualquier ataque.

Del mismo modo, una empresa que tiene menos presencia o está peor posicionada en el entorno digital, puede verse más fácilmente desplazada por narrativas falsas. Irónicamente, una **menor notoriedad pública no necesariamente protege**, sino que hace más frágil la reputación online.

Muchas pequeñas empresas pueden también tener una **falsa sensación de seguridad**, asociando su tamaño a falta de interés para los atacantes.

Sin embargo, si algo nos enseñan las amenazas híbridas es que ninguna marca ni institución, por pequeña que sea, está a salvo de este tipo de ataques, como no lo han estado hasta ahora de las ciberamenazas.

Ahora bien, **ser pequeño no siempre es un inconveniente**. Al tener estructuras más reducidas, las pequeñas y medianas empresas pueden tomar decisiones rápidas y reaccionar de forma más personal y cercana a su comunidad de clientes. Por ejemplo, el dueño de una pyme puede salir inmediatamente en un video en vivo desmintiendo un bulo con tono transparente y cercano, algo más difícil en una multinacional donde los mensajes pasan por varios filtros.

Por eso, aunque las grandes empresas tengan más músculo defensivo, las pymes no están totalmente indefensas ante las amenazas híbridas. Eso sí, deben compensar recursos con vigilancia, proactividad y colaboración, apoyándose quizá en asociaciones sectoriales o instituciones públicas que provean alertas y guía.

Claves para la gestión de la reputación ante amenazas híbridas.

4

Los 5 pilares para la gestión de las amenazas híbridas.

4.1.
Monitoreo y detección.

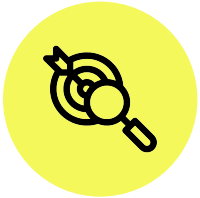
La primera línea de defensa reputacional es la monitorización. Las organizaciones que cuentan con una estrategia de escucha activa y detección temprana pueden reaccionar más rápido, minimizar el daño y tomar el control de la narrativa.

Esta etapa no se limita a instalar una herramienta y esperar alertas: requiere una combinación de tecnología avanzada, criterio humano y, en muchos casos, el apoyo de expertos externos.

El objetivo de esta etapa es detectar de manera temprana contenidos falsificados o sospechosos que circulen en canales digitales y que puedan implicar a la marca, sus productos o sus altos ejecutivos.

4.1.
Monitoreo y detección.

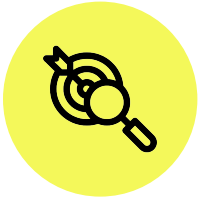
Acciones clave



1. **Escucha activa** y constante en los canales más relevantes para la organización:
    a. Web (noticias, foros, blogs, etc.).
    b. Redes sociales y grupos de mensajería pública ('X', antes Twitter, y otras plataformas con limitaciones de acceso a datos como Meta, YouTube o LinkedIn).
    c. Plataforma publicitaria de Meta (Facebook e Instagram).
    d. Nombres de dominio (*typosquatting*).
    e. Perfiles en redes sociales de altos ejecutivos e identidades digitales vulnerables asociadas a la marca.
2. **Análisis del volumen, tono y contexto** de las conversaciones digitales.
3. **Identificación de anomalías** que puedan indicar la aparición de un contenido manipulado (picos de menciones, cambios en la semántica habitual de las conversaciones).

4.1.
Monitoreo y detección.

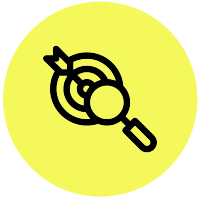
Herramientas



1. **Sistemas de escucha social avanzada** (*social listening*), capaces de rastrear conversaciones públicas en tiempo real en redes sociales, foros, medios digitales, etc. Ejemplos: Brandwatch, Talkwalker, Meltwater, Sprinklr ó Alto Intelligence.
2. **Motores de detección de anomalías por IA**. Identifican cambios repentinos en menciones, emociones o temáticas asociadas a la marca.
3. **Herramientas de detección de *deepfakes***, como Hive Moderation, Deepware Scanner, Microsoft Video Authenticator o Reality Defender. Analizan archivos de audio y vídeo en busca de signos de manipulación generada por IA.
4. **Monitores de *dark web* y canales alternativos.** Rastrean menciones en entornos menos accesibles, como foros cerrados o mensajería semipública.
5. **Supervisión de publicidad programática y plataformas de anuncios**.
6. **Sistemas de vigilancia de dominios y *typosquatting*** para identificar registros sospechosos que simulan la identidad de la marca. Ejemplos: NameGuard y BrandShield.

4.1.
Monitoreo y detección.

Limitaciones y riesgos



- Limitaciones técnicas de **acceso a datos en las plataformas** (Meta, TikTok, LinkedIn). X en menor medida.
- **Opacidad de ciertos canales**, como WhatsApp o Telegram.
- Limitadas **capacidades de monitorización de anuncios fraudulentos** mediante publicidad programática.
- Errores en la detección (**falsos positivos / negativos**) que activan respuestas innecesarias o fallan en alertar.
- **Sobrecarga operativa** por vigilancia 24/7 sin automatización o equipo suficiente. Los grandes ataques se producen en fechas y horarios en los que las compañías tienen menos recursos activados.
- **Dependencia del juicio humano**.
- **Evolución tecnológica constante** que obliga a actualizarse con frecuencia.

4.2.
Evaluación de impacto.

Es el puente entre la simple vigilancia y la activación plena de la gestión de crisis. Confirma la naturaleza y gravedad del incidente, anticipa qué tan lejos podría llegar (en alcance y consecuencias) y organiza las primeras acciones internas de respuesta.

Suele liderar esta fase el equipo de Comunicación, que mide el impacto potencial y coordina el plan de respuesta. Puede recibir el apoyo de Ciberseguridad para la identificación y recopilación de evidencias.

4.2.
Evaluación de impacto.

Acciones clave

1. **Verificación de autenticidad**. ¿Es el contenido realmente falso? A veces lo que parece un *deepfake* puede ser un vídeo real sacado de contexto, o viceversa.
2. **Evaluación de alcance y públicos afectados**. ¿Cuál ha sido la difusión hasta el momento? ¿Está limitado a un nicho de internet? ¿Lo comentan *influencers*, periodistas u otras figuras relevantes? ¿Cuáles son los grupos de interés potencialmente afectados: clientes, empleados, accionistas o inversores, autoridades o reguladores?
3. **Análisis del mensaje y posibles consecuencias**. Cada contenido falso tiene implicaciones diferentes (éticas, legales, comerciales). En esta fase, se empiezan a trazar escenarios: mejor caso (poca difusión, fácil de desmentir) vs. peor caso (viralidad global, impacto duradero).
4. **Consulta de protocolos y plan de crisis existente**. Si la empresa ya cuenta con un plan de crisis reputacional, ¿había contemplado un escenario de *deepfakes*? ¿Qué alineamientos se establecieron (portavoces designados, mensajes preaprobados, procedimientos legales)? Si no existe un plan, se debe seguir una lógica básica de comunicación de crisis: honestidad, rapidez y control de la información.
5. **Notificación interna inicial.** A diferencia de la etapa de “escalada”, aquí basta con involucrar a un núcleo reducido. El director de comunicación debe informar a su superior directo y quizá a uno o dos miembros de la alta dirección más cercanos al tema (por ejemplo, si atañe a un producto específico, al director de esa unidad de negocio). También es oportuno alertar al asesor legal de la empresa y al responsable de Ciberseguridad.

4.2.
Evaluación de impacto.

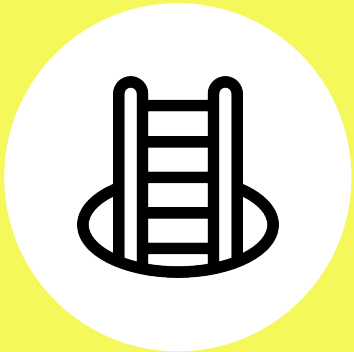
Limitaciones y riesgos

1. **Inexistencia de KPIs específicos** para valorar el impacto real de un *deepfake* en etapas tempranas.
2. Ante la ausencia de datos fiables, **dificultad para dimensionar el daño potencial** y definir el nivel de respuesta necesario.
3. **Decisiones basadas en el criterio profesional** más que en datos cuantificables.
4. **Tiempos de respuesta**. Actuar demasiado pronto puede amplificar el daño; esperar demasiado puede agravarlo.

4.3.
Escalada.

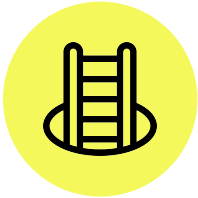
Escalar la situación al resto de la organización significa que deja de ser un asunto manejado por el equipo de Comunicación y se convierte en una prioridad compartida por múltiples departamentos y por la Alta Dirección.

El objetivo de esta etapa es establecer una estructura de responsabilidades y colaboración para afrontar la crisis: asegurar que todas las áreas relevantes (Comunicación, Legal, Recursos Humanos, TI, Seguridad, Operaciones, etc.) estén al tanto, alineadas y trabajando de forma sincronizada bajo un plan común.



4.3.
Escalada.

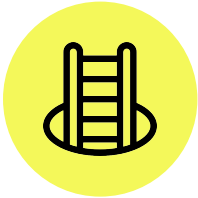
Acciones clave



1. **Reunir al comité de crisis con representantes de Comunicación, Legal, Ciberseguridad, RRHH y Alta Dirección** para confirmar la información conocida, fijar objetivos inmediatos y repartir las tareas urgentes.
2. **Establecer flujos de información seguros y actualizaciones frecuentes**, rompiendo silos. Un mecanismo práctico es crear un canal de comunicación unificado de crisis (por ejemplo, en Slack, Teams o WhatsApp corporativo) donde se comparten novedades al instante y se hacen consultas rápidas. Documentar acuerdos y tareas en un lugar accesible (como un documento colaborativo) ayuda a hacer seguimiento.
3. **Comunicación a la organización**. Dependiendo de la gravedad, puede ser prudente informar al resto de los empleados de que la empresa está lidiando con una crisis de este tipo. Se previene la propagación de rumores internos y se convierte a los empleados en aliados para detectar y desmentir información falsa.
4. **Involucrar a aliados externos**, como agencias de relaciones públicas externas para apoyo adicional, expertos en ciberseguridad para asistencia técnica o para confirmar la falsedad del contenido, así como otros actores industriales, medios de comunicación y verificadores o poderes públicos.

4.3.
Escalada.

Limitaciones y riesgos



- **Fricciones entre departamentos.** Cada área tiene su propia cultura de trabajo y sus prioridades ante una crisis de este tipo. Por ejemplo, Comunicación puede primar la rapidez mientras Legal desea minimizar riesgos legales y recomienda silencio.
- **Lentitud de los procesos de aprobación**. Algunas organizaciones pueden trabarse esperando luz verde formal para cada acción.
- **Comunicaciones internas excesivas**. Existe el riesgo de alarmar a personas que no estaban enteradas del ataque y de que los empleados contribuyan sin querer a la difusión.
- **Comunicación con terceros**. Obliga a lidiar con tiempos y respuestas ajenas. Por ejemplo, Legal puede enviar una solicitud de retirada a una red social pero tendrá que esperar respuesta, mientras Comunicación está ansiosa por que el contenido deje de estar publicado.
- **Agilidad y adaptación**. Un *deepfake* es un evento dinámico. Es posible que mientras la empresa está organizándose internamente, los atacantes o el propio funcionamiento de internet generen giros nuevos.

4.4.
Respuesta.

Busca mitigar el daño reputacional, técnico y operativo derivado de un *deepfake* o contenido fraudulento ya detectado, articulando una respuesta ágil, coordinada y multicanal. La prioridad es restaurar la confianza, proteger a los públicos afectados y evitar que el contenido siga propagándose.

Esta etapa abarca todas las interacciones con el exterior, desde comunicados de prensa, publicaciones en redes sociales, conferencias o vídeos aclaratorios, hasta acciones legales o colaboraciones con autoridades. Incluye también la comunicación con grupos de interés específicos (clientes, socios, empleados, reguladores) para atender sus inquietudes particulares.

4.4.
Respuesta.

Acciones clave

1. **Eliminación o bloqueo del contenido**. Se pueden solicitar la baja de sitios web fraudulentos, eliminar perfiles falsos en redes sociales o servicios digitales, bloquear anuncios engañosos o alertar a los navegadores mediante protocolos de ciberseguridad. Antes de borrar el contenido, es prudente guardar copias seguras con huella temporal (*timestamp*) por si hiciera falta en acciones legales posteriores.
2. **Reportes y escalada a plataformas**. Denuncias formales en Meta, Google, X, etc. En el capítulo 6 te damos más detalles sobre cómo hacerlo.
3. **Comunicación interna.** Hay que informar a los empleados para que no compartan ni se alarmen por el contenido y ofrecerles instrucciones claras sobre cómo responder ante preguntas externas o situaciones incómodas.
4. **Acciones legales o regulatorias.** Reportar a las autoridades o unidades especializadas en delitos tecnológicos y preparar posibles acciones legales si se identifica autoría o se produce un perjuicio económico directo.

4.4. Respuesta.

Acciones clave

5. **Contención y contrarrelato.** Buscamos una comunicación transparente y proactiva, con mensajes directos, sin ambigüedades; con firmeza y serenidad. Implica proporcionar información verificada y asegurar al público que se están tomando medidas efectivas para abordar la crisis, así como actualizar periódicamente si la situación evoluciona.
    - Publicación de mensajes oficiales en canales corporativos: comunicados, notas de prensa, redes sociales, etc. Puede incluir responder a los usuarios que hayan compartido el contenido falso.
    - Contacto con medios de comunicación o verificadores de hechos si el contenido se ha difundido ampliamente. Muchos medios tienen hoy secciones de verificación, por lo que puede ser interesante proporcionarles evidencias para que desmientan el *fake* en sus propias plataformas.
    - Activación de embajadores e *influencers* aliados.
6. **Monitorización.** Durante la respuesta, hay que seguir vigilando en tiempo real cómo está reaccionando la opinión pública a nuestras comunicaciones y adaptar la respuesta.

4.4. Respuesta.

Limitaciones y riesgos

- **Procesos lentos y poco transparentes en plataformas digitales**. No hay protocolos claros de eliminación. Las plataformas no siempre reconocen los *deepfakes* como violaciones de normas, informan de si la denuncia fue aceptada o garantizan que el contenido no reaparezca.
- **Ausencia de regulación efectiva**. La normativa actual es ambigua y no ofrece garantías de actuación rápida ni consecuencias claras.
- **Coordinación compleja**, ya que se trata de un contexto con alta presión y en el que intervienen múltiples actores internos y externos.
- **Riesgo de amplificación**. Una respuesta mal diseñada puede dar más visibilidad al contenido manipulado.

4.5. Capacitación y concienciación.

Superada la fase aguda de la crisis, la gestión no termina: es fundamental aprender de la experiencia y fortalecer a la organización para el futuro mediante capacitación y la introducción de mejoras continuas en los procesos.

El alcance de esta etapa es amplio y de naturaleza preventiva y proactiva. Abarca desde la alta dirección (que puede decidir invertir más en ciertas defensas) hasta el empleado de base (que recibirá orientación para no ser eslabón débil en cadena de desinformación). Incluye también la colaboración con el entorno externo para mejorar colectivamente (industria, reguladores, autoridades).

Esta fase precisa de la colaboración de los equipos de Comunicación Interna y Externa, Ciberseguridad, Experiencia de Usuario, Recursos Humanos y Legal.

4.5. Capacitación y concienciación.

Acciones clave para capacitación

1. **Formación técnica** sobre cómo funcionan este tipo de amenazas, herramientas de detección, etc.
2. **Formación en comunicación de crisis**, como talleres de portavocía bajo presión.
3. **Eventos y sesiones de concienciación presenciales y *online*** para que todas las personas en la empresa sepan identificar contenido falso y entiendan los protocolos de actuación.
4. **Sesiones personalizadas 1-to-1** con miembros del equipo de Ciberseguridad.
5. **Correos, artículos y otras comunicaciones internas** con consejos prácticos sobre *deepfakes* y ciberseguridad. Puede ser interesante crear también contenidos sobre esta materia en sitios webs corporativos dirigidos también a clientes y público general.
6. **Programas de formación y protección específicos para alta dirección.**
7. **Reforzar la cultura corporativa.** Esto puede traducirse en políticas actualizadas o la revisión y reafirmación de los compromisos éticos de la empresa en relación con el uso de la IA.

4.5. Capacitación y concienciación.

Acciones para mejora continua

- Es importante revisar críticamente cómo se gestionó la crisis, identificar aciertos y errores, y **actualizar los protocolos y herramientas** en consecuencia. Se trata de integrar las lecciones aprendidas en el manual corporativo de modo que el sistema de gestión de crisis evolucione.
- Esto también implica **mantenerse al día con la evolución de la tecnología de *deepfakes*** y las estrategias de quienes las usan maliciosamente para ir adaptando la respuesta corporativa a nuevos escenarios.
- Otra herramienta poderosa de mejora es **organizar ejercicios de simulación de crisis** por *deepfakes*.
- También puede ser un buen momento para **compartir la experiencia y construir alianzas externas.** Para la empresa, estos lazos significan que en el futuro tendrá más apoyo y conocimiento a la mano si enfrenta problemas similares.

4.5. Capacitación y concienciación.

Limitaciones y riesgos

- **Entorno dinámico y en rápida evolución**, lo que conlleva un alto riesgo de que la formación que ofrecemos se quede desactualizada.
- **Con el tiempo, la organización puede olvidar lo sucedido y aprendido.**
- **Resistencias internas**, si la crisis reveló errores de gestión y la solución implica cambios —por ejemplo, dar más peso al departamento de Comunicación en decisiones futuras—.
- **Resistencias de *partners* y colaboradores** ante los cambios que pueda ser necesario introducir, como un endurecimiento de las cláusulas contractuales de Ciberseguridad.
- **Limitaciones para la evaluación de la eficacia de la capacitación y ausencia de KPIs específicos**. No se puede esperar a que haya otra crisis para comprobar si ese conocimiento se ha interiorizado.

La importancia de las alianzas para combatir las amenazas híbridas.

5

El reto que plantean las amenazas híbridas es tan complejo que ni siquiera cuando se da en el ámbito empresarial podemos esperar que las compañías lo afronten solas.

De ahí, la importancia de promover la comunicación y la cooperación con otras empresas del mismo sector, pero también de otras industrias, así como con las plataformas y empresas tecnológicas en las que se producen y se desarrollan este tipo de ataques, y los organismos reguladores, academia y sociedad civil.

5.1.
El papel de las plataformas tecnológicas.

Examinar el papel que juegan las grandes plataformas digitales —Meta (Facebook, Instagram, WhatsApp), X (antes Twitter), Google (incluyendo YouTube) y TikTok— en la prevención y gestión de las amenazas híbridas es esencial.

En primer lugar, porque son el principal territorio en el que circulan las campañas de desinformación y *deepfakes*. También porque, por su funcionamiento, favorecen la propagación masiva de contenidos (viralidad) y la difusión de mensajes microsegmentados. Pero, sobre todo, porque **tienen la capacidad, los recursos, la infraestructura y el talento para actuar** en este ámbito y desarrollar, por tanto, soluciones de detección, trazabilidad y moderación de contenidos fraudulentos a gran escala.

En este sentido, la respuesta inicial de las plataformas a la desinformación ha sido autorregulatoria. Cada empresa ha desarrollado políticas internas y normas de funcionamiento de sus comunidades para moderar contenido.

Por ejemplo, Meta (Facebook) implementó desde 2016-2018 un amplio programa voluntario de verificación de datos tras el escándalo de Cambridge Analytica, colaborando con terceros. Para 2023, este programa abarcaba contenidos en más de 60 idiomas con ayuda de unas 100 organizaciones externas de *fact-checking*. Del mismo modo, se han adherido a códigos de conducta voluntarios. En 2018, plataformas como Facebook, Google y Twitter suscribieron el Código de Buenas Prácticas contra la Desinformación de la UE, al que luego se sumaron Microsoft y TikTok.

El problema es que **la autorregulación de las grandes plataformas ha demostrado ser insuficiente**, generando espacios de indefensión para ciudadanos, empresas e instituciones que quedan en situación de desprotección frente a los grandes actores tecnológicos. De hecho, desde 2022, han sido también sonadas las noticias sobre los despidos y recortes de personal en los equipos de contenidos y *fact-checking* de algunas de estas plataformas, algo que podría debilitar su capacidad de reacción ante amenazas híbridas y que ha motivado una vigilancia más intensa por parte de las autoridades, la sociedad civil y los medios de comunicación.

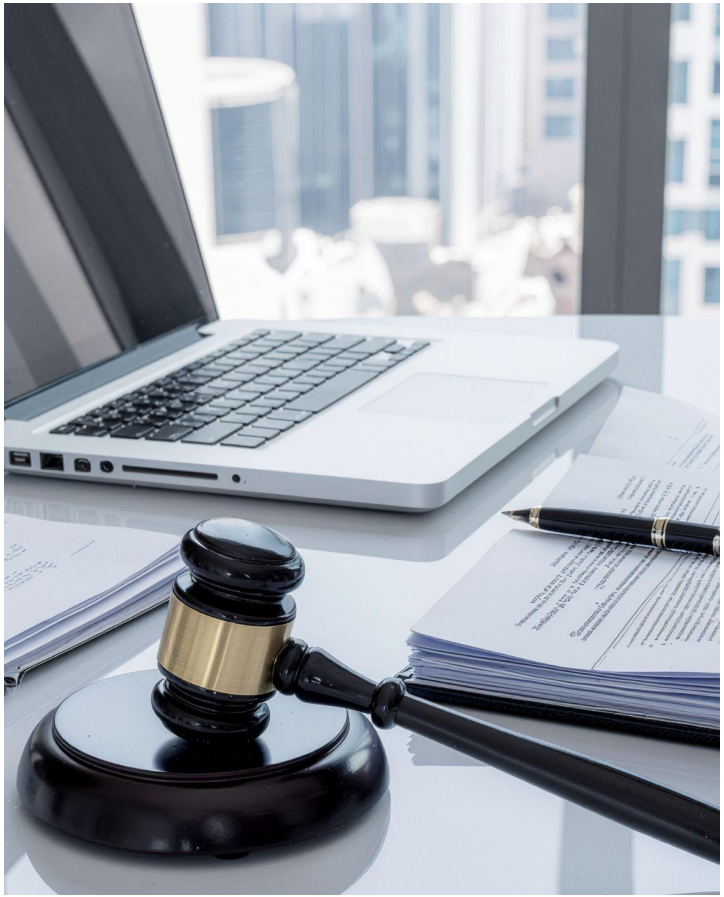
5.2.
Marcos legales como herramientas de protección.

La ineficacia de la autorregulación hace necesario el impulso de marcos legales como herramientas de protección.

La Unión Europea o el gobierno del Estado de California, entre otros, han implementado **normativas clave para mitigar los riesgos de *deepfakes*, ciberataques y desinformación**, buscando equilibrar el ecosistema digital y fortalecer la protección de los derechos fundamentales. Asimismo, han establecido **estándares** de transparencia, responsabilidad y trazabilidad en el uso de tecnologías avanzadas.

En este sentido, es especialmente relevante la DSA (*Digital Services Act*) de la Unión Europea que impone obligaciones específicas a las plataformas digitales para garantizar un uso responsable y transparente de la IA, con especial mención a los contenidos generados o manipulados mediante técnicas de *deepfakes*. La DSA exige que los contenidos sintéticos que puedan inducir a error (como imágenes, vídeos y audios que imiten a personas reales) estén claramente etiquetados. Obliga también a las plataformas a disponer de mecanismos eficaces para su detección, notificación y retirada cuando puedan causar daño.

En este marco de las políticas públicas, **las empresas han de adoptar un papel proactivo** para dar seguimiento a los procesos regulatorios, presentar las denuncias oportunas y exigir marcos regulatorios apropiados.



Solo cuando el ecosistema **actúa de forma integrada**, se eleva el nivel de protección y respuesta a las amenazas híbridas para todos y todas.

5.3.
Espacios de diálogo y alianzas.

Los foros internacionales y las iniciativas de alianzas público-privadas facilitan la identificación de patrones de riesgo, la compartición de inteligencia y la creación de respuestas coordinadas.

En febrero de 2024, durante la Munich Security Conference, las grandes tecnológicas se comprometieron a combatir los *deepfakes* electorales, con herramientas compartidas, transparencia y campañas de concienciación.

Si buscamos ejemplos de cooperación intersectorial, ya existen alianzas entre bancos y empresas tecnológicas para desarrollar algoritmos antifraude y acuerdos entre plataformas digitales y medios. Es el caso de la Coalition for Content Provenance and Authenticity (C2PA), que demuestra el impacto positivo de la cooperación.*

Por otro lado, una respuesta eficaz requiere una **acción conjunta multilateral**, en la que también pueden y deben entrar las universidades y centros de investigación (para ayudar en el desarrollo de medidas detección o la definición de estándares) y la sociedad civil (con funciones de vigilancia pública y alfabetización digital).

¿Cómo actuar en caso de un ataque contra tu compañía?

6

6.1.
Cómo responder a las amenazas híbridas.

Ante la creciente amenaza de ataques que utilizan contenidos manipulados, es clave que las organizaciones establezcan un **protocolo interno claro para identificar, documentar y responder a este tipo de incidentes**.

Este procedimiento debe facilitar la coordinación entre los distintos departamentos implicados en la protección de la reputación corporativa, asegurando una respuesta ágil y eficaz.

Este protocolo debe incluir **instrucciones específicas para la persona que detecte el ataque**, así como para el equipo encargado de su gestión.

Establecer este tipo de procedimientos no solo permite actuar con rapidez ante una crisis, sino que también refuerza la cultura de vigilancia y prevención en la organización.

A continuación, se detallan algunas recomendaciones generales y los mecanismos de denuncia disponibles en plataformas como X, Meta y YouTube.

6.2.
Primeros pasos para denunciar un ataque.

Recomendaciones generales y comunes a todas las plataformas digitales.

Realiza una captura de pantalla de la publicación o anuncio.

Copia la URL del post o anuncio.

Recoge esta información y realiza la documentación de los casos detectados.

Reporta SIEMPRE la publicación a través de la plataforma de redes sociales.

6.3. Mecanismos de denuncia por plataforma.

	Google - Youtube	Meta	X
PASOS	1. Haz click en los 3 puntos verticales en la derecha del anuncio y procede a 'Denunciar anuncio' o 'Report ad'. 2. Marca "It violates Google policies". 3. Marca "It's misleading or a scam". 4. En este punto, debería aparecer automáticamente el Link para el ad o listing. 5. Añade tus comentarios detallando la razón por la que denuncias el anuncio.	**En el momento de visualizar el anuncio:** 1. En el feed, haz clic en *** junto al anuncio. 2. Marca "Denunciar anuncio" y sigue las instrucciones. **Más tarde:** 1. Ve a la Biblioteca de Anuncios de Meta 2. En "Buscar anuncios", localiza el contenido que quieres denunciar a través de palabras clave y sigue los puntos 1 y 2.	1. Sitúate sobre el post y selecciona *** (localizado en el zona superior del mismo). 2. Marcar "Report post" y seguir las instrucciones.
RECURSOS ÚTILES	**Políticas de Google / Youtube** **Directrices de privacidad** **Canal de denuncia** de contenidos inapropiados	**Normas de publicidad de Meta** **Biblioteca de anuncios** **Información sobre fraudes, estafas y prácticas engañosas**	**Información de X para reportar posts, listas y otras circunstancias** **Información de X para incidencias de autenticidad**

6.4.
¿Qué pasa después?

Una vez has denunciado el contenido fraudulento en la plataforma, solo queda esperar la respuesta de la misma.

La Ley de Servicios Digitales (Digital Services Act) no establece plazos concretos para responder a reportes de contenido específico, pero obliga a aquellas plataformas muy grandes (*very large online platforms*, con más de 45 millones de usuarios) a implantar sistemas de moderación de contenido más fiables y transparentes. Esta normativa busca que las plataformas indiquen cómo gestionan reclamaciones sin imponer tiempos exactos de actuación.

Tampoco existen estimaciones de cuánto tiempo suelen demorarse esas respuestas, salvo por algún estudio *ad hoc*, como el que evaluó la rapidez con la que la plataforma X retira imágenes íntimas no consentidas (NCIM), incluyendo *deepfakes*. Los resultados mostraron que cuando el reporte se realiza por infracción de derechos de autor se eliminan las imágenes en un plazo de hasta 25 horas. En otros casos, este pueden llegar a superar las tres semanas.

6.5. Denuncia ante las autoridades.

El INCIBE (Instituto Nacional de Ciberseguridad) propone seguir el siguiente procedimiento para informar a las autoridades de un posible delito de suplantación de identidad:

1

Contactar con el centro CERT de INCIBE y reportar el caso a través de la línea gratuita de Ayuda en Ciberseguridad 017 y/o de los canales de mensajería instantánea WhatsApp 900 116 117 y Telegram @INCIBE017. También existe un formulario, disponible en:

https://www.incibe.es/incibe-cert/incidentes/notificaciones

2

Formular la correspondiente denuncia ante las fuerzas y cuerpos de seguridad del Estado y/o ante el Juzgado de Guardia correspondiente.

3

Informar al Observatorio Español de Delitos Informáticos a través del siguiente formulario:

https://oedi.es/formulario-fake-news/

Esta alerta se envía con fines estadísticos; no sustituye a la denuncia formal.

4

Si el *deepfake* ha ocasionado perjuicios económicos o legales (fraudes, declaraciones u operaciones que puedan generar sanciones posteriores), es recomendable renovar los documentos comprometidos e informar a la AEAT (Agencia Tributaria) y al Banco de España (en el caso de que se tema un uso indebido de documentación para solicitar préstamos o generar operaciones financieras).

Cuestionario de autoevaluación.

7

¿Está tu organización preparada para sobrevivir en este nuevo escenario de amenazas híbridas?

A continuación te proponemos una serie de preguntas que te ayudarán a reflexionar sobre el nivel de preparación de tu empresa frente a este nuevo tipo de desafíos reputacionales, técnicos y organizativos.

No se trata de tener todas las respuestas, sino de iniciar conversaciones clave para anticiparse mejor y responder con agilidad cuando sea necesario.

Autoevaluación.

Foro-ia

7.1. Preparación organizativa.

- ¿Existe un protocolo claro de actuación ante contenidos manipulados (vídeos, audios, mensajes en redes, suplantaciones de identidad)?
- ¿Habéis definido quién lidera la gestión de este tipo de crisis dentro de la organización?
- ¿Existe un equipo o comité multidisciplinar (Comunicación, Ciberseguridad, Legal, Dirección, etc.) preparado para activarse en estos casos?
- ¿Se evalúa periódicamente la exposición y vulnerabilidades de la organización ante estas amenazas?

7.2. Personas y cultura.

- ¿Se ha formado a portavoces y directivos para responder de manera adecuada a situaciones de crisis provocadas por ataques con *deepfakes* o campañas de desinformación?
- ¿Están los empleados capacitados para identificar contenido falso y saber cómo actuar si lo detectan?
- ¿Se fomenta una cultura organizativa que favorezca la cooperación ágil entre funciones clave en ese contexto?

7.3. Tecnología y monitorización.

- ¿Se dispone de sistemas que alerten de menciones anómalas, suplantaciones de identidad o contenidos sospechosos?
- ¿Se supervisan canales relevantes como redes sociales, plataformas publicitarias, mensajería semiprivada y foros cerrados?
- ¿Existen mecanismos internos para escalar rápidamente cualquier amenaza que detecte un empleado o colaborador?

Autoevaluación.

Foro-ia

7.4. Comunicación y confianza.

- ¿Se informa de forma proactiva a los clientes sobre fraudes emergentes que puedan afectarles (ej. llamadas fraudulentas, emails falsos, etc.)?
- ¿Se dispone de mensajes claros, canales preparados y portavoces entrenados para responder públicamente de forma ágil y creíble?
- ¿Está la organización preparada para articular una respuesta multicanal que combine comunicación pública, atención directa y retirada de contenidos falsos?

7.5. Ecosistema y colaboración externa.

- ¿Tienen los proveedores y aliados protocolos de actuación si son víctimas o canales involuntarios de un ataque?
- ¿Se incluye la preparación ante amenazas híbridas como parte de la estrategia de relación con terceros (evaluación de riesgos, contratos, auditorías)?
- ¿Hay en marcha colaboraciones con otras empresas del sector o asociaciones para compartir buenas prácticas, alertas o aprendizajes?
- ¿Se mantiene una relación directa con plataformas digitales (como redes sociales o motores de búsqueda) para actuar con rapidez ante contenido falso grave?

Autoevaluación.

Foro-ia

7.6. Colaboración institucional y fuerzas de seguridad.

- ¿Se tienen identificados los canales de contacto con cuerpos policiales, unidades de delitos tecnológicos u otras autoridades relevantes en caso de incidente?
- ¿Se tiene acceso a los recursos, guías o programas de formación que ofrecen las fuerzas y cuerpos de seguridad del Estado sobre fraude, desinformación y ciberataques?
- ¿Se han establecido vínculos con organismos públicos, centros de ciberseguridad o instituciones que puedan asesorarnos o actuar con rapidez en una situación de crisis?

Bibliografía.

8

Fuentes y referencias bibliográficas.

Alto Intelligence. *Hybrid Threats and the Amplifying Power of AI: Five Strategic Scenarios. 2025.* *https://www.altointelligence.com/wp-content/uploads/2025/03/Hybrid-Threats-and-the-Amplifying-Power-of-AI_Alto_Intelligence.pdf*

Centro Europeo de Excelencia para la Lucha contra las Amenazas Híbridas (Hybrid CoE). *Hybrid threats.* https://www.hybridcoe.fi/hybrid-threats/

Comisión Europea y Centro Europeo de Excelencia para la Lucha contra las Amenazas Híbridas (Hybrid CoE). *The Landscape of Hybrid Threats: A Conceptual Model. Public Version.* 2021. https://publications.jrc.ec.europa.eu/repository/handle/JRC123305

Mazzucchi, N. *AI-based technologies in hybrid conflict: The future of influence operations.* Hybrid CoE Paper 14. Junio 2022. https://www.hybridcoe.fi/publications/hybrid-coe-paper-14-ai-based-technologies-in-hybrid-conflict-the-future-of-influence-operations/

Consejo de la Unión Europea. *Amenazas híbridas.* https://www.consilium.europa.eu/es/policies/hybrid-threats/

ISMS Forum. *Deepfakes: riesgos, casos reales y desafíos en la era de la IA.* Marzo 2025. https://www.ismsforum.es/ficheros/descargas/deepfake-final1742458135.pdf

INCIBE– Instituto Nacional de Ciberseguridad. *Historias reales: el deepfake de mi jefe circulando en la red.* https://www.incibe.es/empresas/blog/historias-reales-deepfake-mi-jefe-circulando-red

Washington, D.C.: U.S. Department of Homeland Security. *Increasing Threat of Deepfake Identities.* 2019. https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf

Washington, D.C.: National Security Agency, FBI, and Cybersecurity and Infrastructure Security Agency. *Contextualizing Deepfake Threats to Organizations.* Cybersecurity Information Sheet. Septiembre 2023. https://media.defense.gov/2023/Sep/12/2003298925/-1/-1/0/CSI-DEEPFAKE-THREATS.PDF

Luu, Alice. "Deepfake & Misinformation Rapid Response Framework for Enterprise Communications Teams". *Insights – Manhattan Strategies.* 29 mayo 2025. https://www.manhattanstrategies.com/insights/deepfake-misinformation-rapid-response-playbook#:~:text=5

Regula Forensics. *Deepfake Trends 2024.* 2024. https://regulaforensics.com/resources/deepfake-trends-2024-report/

KPMG y Reality Defender. *Deepfakes: Real Threat.* 2023. https://kpmg.com/kpmg-us/content/dam/kpmg/pdf/2023/deepfakes-real-threat.pdf

Global Market Estimates. *Deepfake Detection Software Market.* Enero 2025. https://www.globalmarketestimates.com/market-report/deepfake-detection-software-market-4420

Fuentes y referencias bibliográficas.

Ortíz, Mariano. “Amenazas híbridas contra empresas”. *Tarlogic. Blog.* 14 abril 2020. https://www.tarlogic.com/es/blog/amenazas-hibridas-contra-empresas/

Baruchin, R. “Disinformation Doesn’t Care About Your Size: Protecting Brand Reputation for SMBs”. *Cyabra. Blog.* 9 junio 2025. https://cyabra.com/blog/disinformation-doesnt-care-about-your-size-protecting-brand-reputation-for-smbs/

Mut-Camacho, M. “Las empresas ante la desinformación. La necesidad de un nuevo enfoque metodológico”. Vivat Academia Revista de Comunicación 155:113-129. 2022. DOI:10.15178/va.2022.155.e1327 https://www.vivatacademia.net/index.php/vivat/article/view/1327

European Commission. *The 2018 Code of Practice on Disinformation*. [en línea] Digital Strategy (estrategia digital de la UE), 2018. Disponible en: https://digital-strategy.ec.europa.eu/en/library/2018-code-practice-disinformation

Deloitte. *Generative AI and the fight for trust*. Deloitte. 2024. https://www.deloitte.com/content/dam/assets-zone3/us/en/docs/services/consulting/2025/generative-ai-and-the-fight-for-trust.pdf

World Economic Forum. *Global Risks Report 2025*. Enero 2025. https://www.weforum.org/publications/global-risks-report-2025/

Edelman. *2025 Edelman Trust Barometer. Trust and the crisis of grievance.* 2025. https://www.edelman.com/trust/2025/trust-barometer

Bueermann G. y Perucica N. “How can we combat the worrying rise in the use of deepfakes in cybercrime?”. *World Economic Forum*. 19 mayo 2023. https://www.weforum.org/stories/2023/05/how-can-we-combat-the-worrying-rise-in-deepfake-content/

Hogdson, C. “World’s biggest tech companies pledge to fight AI-created election ‘deepfakes’”. *Financial Times*. 16 febrero 2024. https://www.ft.com/content/8dcbc162-a0f5-47ce-a2ea-5fb25cb160c5

Sahuquillo, M.R. “La guerra híbrida de la desinformación en la UE alcanza un nivel sin precedentes”. *El País*. 24 noviembre 2024. https://elpais.com/tecnologia/2024-11-24/la-guerra-hibrida-de-la-desinformacion-en-la-ue-alcanza-un-nivel-sin-precedentes.html

Feiner, Lauren. “Trump signs the Take It Down Act into law”. *The Verge*. 29 mayo 2025. https://www.theverge.com/news/661230/trump-signs-take-it-down-act-ai-deepfakes

González, F. “Debemos proteger nuestra capacidad de reconocer a seres humanos reales”: más de 400 expertos firman carta contra los deepfakes. *Wired España*. 21 febrero 2024. https://es.wired.com/articulos/mas-de-400-expertos-firman-carta-contra-los-deepfakes

Ramírez Quesada, Aurora. “La inteligencia artificial, una aliada inesperada contra los ‘deepfakes’”. *Agencia SINC*. 10 de abril de 2025.

Fuentes y referencias bibliográficas.

Galleti, S. y Pani, M. “How Ferrari Hit the Brakes on a Deepfake CEO.” *MIT Sloan Management Review*. 27 enero 2025. https://sloanreview.mit.edu/article/how-ferrari-hit-the-brakes-on-a-deepfake-ceo/

Amantegui Guezala, A. “Roban 35 millones de dólares clonando la voz de un director con inteligencia artificial”. *La Vanguardia*, 16 octubre 2021. https://www.lavanguardia.com/tecnologia/20211016/7794770/roban-35-millones-dolares-clonando-voz-director-inteligencia-artificial-pmv.html

Hudson, C. “Scammers Use AI-Generated Voice Deepfake to Steal Almost $250,000 from CEO”. IdentityIQ. 6 septiembre 2019. https://www.identityiq.com/articles/scammers-use-ai-generated-voice-deepfake-to-steal-almost-250000-from-ceo

Schofield, H. “La sofisticada estafa del falso ministro de Francia con máscara de goma que se robó US$90 millones”. *BBC News*. 20 junio 2019. https://www.bbc.com/mundo/noticias-48711980

Backovsky, D. “Exploring hybrid threats with Alex Romero”. Podcast. *Berlin Security Beat*. 13 sept 2024. https://open.spotify.com/episode/30ed0xuVVDSc6sz9p5KBbV

Newtral. “Mayor lucha contra la desinformación y más transparencia: las obligaciones que tienen que cumplir las plataformas con la DSA”. *Newtral*. 17 agosto 2023. https://www.newtral.es/digital-services-act-desinformacion/20230817/

Coalition for Content Provenance and Authenticity (C2PA). https://c2pa.org/

Caso RTVE.

Pena, P. “Falso vídeo suplanta a Ana Blanco con proyecto Amancio Ortega”. *Verifica RTVE*. 14 diciembre 2023. https://www.rtve.es/noticias/20231214/falso-video-suplanta-ana-blanco-proyecto-amancio-ortega/2467023.shtml

Marco, S. y Navarro, M. “Falso vídeo RTVE hablando de proyecto financiero Ortega-Amancio”. *Verifica RTVE Noticias*. 22 marzo 2024. https://www.rtve.es/noticias/20240322/falso-video-rtve-hablando-proyecto-financiero-ortega-amancio/16028148.shtml

Marco, S. “Falso RTVE no promociona proyecto financiero Marta Ortega deepfake”. *Verifica RTVE*. 30 septiembre 2024. https://www.rtve.es/noticias/20240930/falso-rtve-no-promociona-proyecto-financiero-marta-ortega-deepfake/16268287.shtml

Fuentes y referencias bibliográficas.

Caso ARUP.

Magramo, K. “British engineering giant Arup revealed as $25 million deepfake scam victim”. *CNN*. 17 mayo 2024. https://edition.cnn.com/2024/05/16/tech/arup-deepfake-scam-loss-hong-kong-intl-hnk

Caso Pentágono.

Kardoudi, O. “El Pentágono víctima de fake con inteligencia artificial”. *El Confidencial / Novaceno*. 23 mayo 2023. https://www.elconfidencial.com/tecnologia/novaceno/2023-05-23/pentagono-fake-inteligencia-arficial_3634851/

Caso Qantas.

Taylor, J. “Qantas attack reveals one phone call is all it takes to crack cybersecurity’s weakest link: humans.” *The Guardian*. 5 julio 2025. https://www.theguardian.com/business/2025/jul/06/qantas-attack-reveals-one-phone-call-is-all-it-takes-to-crack-cybersecuritys-weakest-link-humans

Recursos para reportar contenidos fraudulentos en plataformas

GOOGLE. *Política sobre contenido manipulado de forma técnica*. Ayuda de YouTube. https://support.google.com/youtube/answer/10684207?hl=es-419

GOOGLE. *Cómo informar sobre un contenido inapropiado*. Ayuda de YouTube. https://support.google.com/youtube/answer/2801895?hl=es-419

GOOGLE. *Normas de la comunidad de YouTube*. Ayuda de YouTube. https://support.google.com/youtube/answer/2802027?hl=es

META. *Advertising Standards*. Meta Transparency Center. https://transparency.meta.com/policies/ad-standards/

META. *Biblioteca de anuncios: anuncios políticos y sobre temas sociales en España*. Facebook Ads Library. https://www.facebook.com/ads/library/?active_status=active&ad_type=political_and_issue_ads&country=ES

META. *Prácticas engañosas y fraudes – Normas de anuncios*. Meta Transparency Center. https://transparency.meta.com/es-es/policies/ad-standards/fraud-scams/fraud-scams-deceptive-practices/

X. *Cómo denunciar una publicación*. Centro de ayuda de X. https://help.x.com/en/safety-and-security/report-a-post

X . *Política sobre autenticidad*. Centro de ayuda de X. https://help.x.com/en/rules-and-policies/authenticity

Sobre el Foro IA.

El Foro IA en Marketing, Comunicación y Experiencia de Cliente (MCX) nace con el objetivo de generar un espacio para aprender en comunidad, construir conocimiento y compartir recomendaciones y pautas que ayuden a los profesionales MXC a navegar por el tsunami que representa la IA Generativa.

Creemos en la comunidad, en la capacidad de establecer conexiones relevantes y en el debate como herramienta de transformación. Nuestro enfoque abarca tres áreas de trabajo fundamentales:

Alfabetización

Impulsar una mejor comprensión de la tecnología subyacente, aportando una perspectiva accesible para todos y sin necesidad de ser tecnólogos.

Aplicación funcional

Integrar de manera efectiva las nuevas IAs generativas en el entorno del marketing, la comunicación y la experiencia de cliente, reimaginando el "martech stack" del futuro.

Impacto social

Poner foco, como profesionales comprometidos con el impacto positivo de las empresas en la sociedad, en los aspectos éticos de la IA y el avance de la regulación.

Si te interesa participar y recibir información sobre las actividades del Foro, déjanos tu email

Foro-ia