GUÍA RED TEAM · SOLUTECH

El Lado Oscuro de la Inteligencia Artificial

Cómo te atacan con IA en 2026 —deepfakes, fraude del CEO, herramientas de IA maliciosa— y qué exige ahora el Reglamento Europeo de IA. Guía práctica para pymes de Almería y Murcia.

Elaborado por **Xavi Alonso — Solutech**
solutech.blog · Agosto 2026

Índice

Introducción

El 15 de febrero de 2026, la Guardia Civil tuvo que salir a alertar públicamente de perfiles falsos generados con IA que suplantaban a sus propios agentes en redes sociales. Ese mismo mes, Google retiró un anuncio fraudulento que usaba la imagen y la voz de Pablo Motos para promocionar una inversión ficticia. Y unas semanas antes, el CFO de una empresa de logística española autorizó una transferencia de 580.000€ tras recibir una videollamada de su CEO: rostro, voz, gestos, tono, todo idéntico. Tres horas después descubrieron que el CEO real estaba en un vuelo a Tokio, sin conexión. La videollamada era un deepfake.

No son casos aislados ni ciencia ficción. Son el nuevo terreno de juego de 2026, y esta guía trata precisamente de eso: la misma tecnología que en nuestra guía anterior usábamos para ahorrar horas de trabajo en tu empresa, en manos equivocadas, se convierte en el arma de fraude corporativo de mayor crecimiento del mundo.

Esta no es una guía para asustarte sin más. Es una guía Red Team: te vamos a explicar cómo piensa y actúa el atacante, con datos verificados y casos reales, para que sepas exactamente dónde reforzar tu empresa. Y como el 2 de agosto de 2026 ha entrado en vigor la mayor parte del Reglamento Europeo de IA, también te explicamos qué obligaciones legales nuevas te afectan ya, aunque tu empresa nunca haya tenido intención de hacer nada malo con la IA.

Los datos que vertebran esta guía:

- **23% de las empresas españolas** ya han sufrido un intento de fraude con deepfake, con una pérdida media de **600.000€** cuando el ataque tiene éxito.
- **929 millones de euros** en fraudes deepfake a nivel global en 2025, el triple que en 2024.
- **122.223 ciberincidentes** gestionados por INCIBE en España en 2025, un 26% más que el año anterior.
- El ransomware se **duplicó** en España en 2025 (176 → 392 casos), con un rescate medio de **80.000-150.000€** por incidente.
- Desde el **2 de agosto de 2026**, cualquier empresa que use IA generativa tiene nuevas obligaciones legales de transparencia, con sanciones de hasta 35 millones de euros o el 7% de la facturación global para los incumplimientos más graves.

El nuevo arsenal del ciberdelincuente

Durante años, el perfil del atacante que amenazaba a una pyme era relativamente predecible: phishing genérico, malware conocido, fuerza bruta contra contraseñas débiles. La IA generativa ha cambiado las reglas en tres frentes concretos.

1. Personalización a escala industrial

Antes, un email de phishing convincente y personalizado requería horas de investigación manual sobre la víctima. Hoy, un atacante hace scraping automático de LinkedIn, entrevistas en YouTube y podcasts empresariales para construir un perfil completo de un directivo en minutos, y genera con IA un mensaje adaptado a su tono, su entorno y su situación real. El resultado: campañas de fraude hasta **4,5 veces más efectivas** que el phishing tradicional genérico.

2. Contenido sintético indistinguible

Clonar una voz ya solo requiere **3 segundos de audio**, con hasta un 85% de fidelidad. Los directivos que aparecen en LinkedIn, ruedas de prensa o vídeos corporativos están proporcionando ese material de entrenamiento de forma gratuita y masiva, sin saberlo. Los deepfakes de vídeo en tiempo real ya operan con una latencia inferior a 200 milisegundos, lo bastante rápido para sostener una videollamada en directo sin que se note el retardo.

3. Herramientas de IA sin ninguna restricción ética

Mientras que ChatGPT, Claude o Gemini incorporan salvaguardas que impiden generar malware o campañas de phishing, en la dark web circulan desde hace tiempo "gemelos malvados" de estos modelos, diseñados explícitamente para uso criminal, sin ningún tipo de filtro.

Deepfakes: la amenaza que ya cuesta 600.000€ por incidente

Un deepfake es contenido audiovisual sintético —vídeo, audio o imagen— generado o modificado con IA para hacer parecer real algo que no lo es. Lo que hace peligrosos a los deepfakes de 2026 no es solo su realismo, sino su disponibilidad: las plataformas capaces de generarlos son en muchos casos gratuitas, no requieren conocimientos técnicos y permiten un uso anónimo.

Los tipos de fraude con deepfake más habituales en empresas

- **Suplantación de directivo (CEO/CFO fraud):** una videollamada o audio "urgente" de un superior autorizando una transferencia o una decisión fuera de proceso.
- **Suplantación de proveedor:** un cambio de cuenta bancaria comunicado supuestamente por un proveedor habitual, con voz o vídeo clonados para reforzar la credibilidad ante una llamada de verificación.
- **Vishing con voz clonada:** llamadas telefónicas con la voz de un familiar, jefe o compañero, generadas a partir de segundos de audio público.
- **Perfiles falsos en redes sociales:** identidades sintéticas completas (foto, biografía, publicaciones) usadas para ganarse la confianza antes de pedir dinero o datos.

Por qué las defensas tradicionales ya no bastan

Gartner prevé que, para 2026, **el 30% de las empresas considerará que las soluciones de verificación de identidad basadas solo en contenido no son fiables**. Dicho de otra forma: mirar si "la cara se ve rara" o si "la voz suena metálica" deja de ser una defensa válida a medida que mejora la calidad de generación. La detección fiable ya no puede depender de identificar fallos visuales o sonoros; tiene que basarse en el **proceso de verificación**, no en el contenido en sí. Es la idea central del protocolo que verás más adelante en esta guía.

Anatomía de un ataque: así funciona el fraude del CEO

Para defenderte de algo, conviene entender su estructura, sin entrar en detalles técnicos operativos que faciliten su ejecución. Un fraude de suplantación de directivo con IA sigue, de forma consistente, una misma secuencia:

Fase 1 — Reconocimiento. El atacante recopila material público del directivo a suplantar: apariciones en vídeo, entrevistas, notas de voz, publicaciones en LinkedIn. Cuanto más expuesto está un directivo en medios y redes, más fácil resulta esta fase.

Fase 2 — Generación del señuelo. Con ese material, se genera el contenido sintético: un audio, un vídeo o ambos, suficientemente convincente para el canal de contacto elegido.

Fase 3 — Creación de urgencia. El mensaje siempre incluye un elemento de presión temporal y confidencialidad: "es urgente", "no se lo comentes a nadie más", "necesito esto antes de que cierre el banco". La urgencia y el secretismo son, con diferencia, la señal de alerta más fiable en todo el proceso.

Fase 4 — Solicitud fuera de proceso. La petición final —una transferencia, un cambio de datos bancarios, un acceso— se pide saltándose el procedimiento habitual de aprobación, apelando a la autoridad de quien parece estar pidiéndolo.

La conclusión práctica: el punto en el que se puede frenar este ataque no es detectando si el vídeo es falso, sino comprobando si la petición sigue el proceso normal de tu empresa. Si no lo sigue, no importa lo convincente que parezca: se detiene y se verifica por un canal alternativo ya conocido.

IA maliciosa: WormGPT, FraudGPT y los "gemelos malvados" de ChatGPT

Desde 2023 circulan en foros de la dark web y canales de Telegram versiones de modelos de lenguaje sin ningún filtro de seguridad, comercializadas abiertamente por suscripción, con precios que van desde unos 60€ hasta varios miles de euros para configuraciones privadas. Los nombres más conocidos —WormGPT, FraudGPT y sus sucesores— se promocionan explícitamente para:

- Redactar correos de phishing y páginas web fraudulentas indistinguibles de las legítimas.
- Generar código malicioso y malware personalizado.
- Facilitar el compromiso de cuentas en redes sociales y mensajería.
- Preparar campañas de compromiso de correo corporativo (BEC).

Lo relevante para una pyme no es memorizar estos nombres, sino entender la implicación de fondo: **la barrera técnica para lanzar un ataque sofisticado ha desaparecido.** Ya no hace falta ser un experto en programación para generar malware funcional o una campaña de phishing gramaticalmente perfecta y contextualizada. Esto explica en buena parte por qué el volumen de incidentes ha subido un 26% en España en un solo año.

Ransomware potenciado por IA: los números de 2026

El ransomware sigue siendo, según INCIBE y ENISA, la amenaza de mayor impacto económico para las empresas españolas, y la IA está acelerando su ciclo de ataque en dos puntos concretos: la identificación automatizada de objetivos vulnerables (independientemente de su tamaño) y la redacción de las comunicaciones de extorsión, ahora más personalizadas y con mayor presión psicológica.

Los datos de INCIBE para 2025 en España:

Indicador	Dato 2025	Variación
Incidentes de ciberseguridad totales	122.223	+26% vs 2024
Ataques de ransomware	392	+122% vs 2024 (176 casos)
Rescate medio a empresa mediana	80.000€ - 150.000€	—
Tiempo medio hasta el cifrado	Menos de 5 días	Antes, semanas
Coste total medio por incidente (UE)	+1,6 millones de euros	Incluye inactividad y recuperación

El modelo dominante ya no es "cifrar y pedir rescate": es **doble extorsión**. Primero se exfiltran los datos, después se cifran los sistemas. Pagar el rescate no elimina el riesgo de que esos datos ya robados se publiquen o se vendan igualmente. Por eso la recomendación oficial de INCIBE sigue siendo no pagar nunca el rescate: no garantiza recuperar la información y confirma a los atacantes que tu empresa es un objetivo rentable.

Casos recientes con nombre propio en 2026 (Inditex, Endesa) muestran además un patrón relevante para cualquier pyme proveedora: el origen del incidente estuvo en un **proveedor tecnológico externo** con acceso a los sistemas. Si tu empresa trabaja como proveedora de cuentas más grandes, tu nivel de seguridad ya no es solo un asunto interno: es parte de la cadena de suministro que tus clientes tienen que auditar.

El nuevo marco legal: Reglamento Europeo de IA (AI Act) desde agosto de 2026

El **2 de agosto de 2026** ha entrado en aplicación la mayor parte de las obligaciones de transparencia del Reglamento (UE) 2024/1689, conocido como AI Act. No es una norma dirigida solo a grandes tecnológicas: afecta a **cualquier organización que desarrolle, comercialice o utilice sistemas de IA en el mercado europeo**, con independencia de su tamaño.

Qué obliga exactamente el artículo 50

El eje central de esta nueva fase es el **artículo 50**, que establece obligaciones de transparencia en dos planos:

1. **Interacción con sistemas de IA.** Si tu empresa usa un chatbot o asistente automatizado de cara al cliente, debe quedar claro que la persona está interactuando con una IA, salvo que resulte obvio por el contexto.
2. **Contenido sintético.** Si tu empresa genera o publica contenido de audio, imagen, vídeo o texto creado o modificado sustancialmente con IA, debe informarse de que es contenido sintético cuando pueda inducir a error sobre su autenticidad. Los proveedores de estas herramientas deben, además, marcar sus resultados en un formato legible por máquina que permita detectar que han sido generados artificialmente.

Un matiz importante: la obligación **no es absoluta**. No hay que etiquetar cualquier foto retocada, cualquier texto corregido con IA o cualquier uso auxiliar y menor de estas herramientas. El criterio es si la aportación de la IA es sustancial y puede confundir a una persona razonable sobre el origen del contenido.

Qué debe hacer tu empresa antes de que te lo pidan

- **Inventariar** qué herramientas de IA generativa usa tu empresa hoy (redacción de contenido, chatbots, generación de imágenes, etc.).
- **Diferenciar** contenido totalmente sintético de contenido parcialmente asistido por IA.
- **Establecer quién valida** cada publicación que use IA generativa antes de que salga de la empresa.
- **Coordinarte con proveedores externos** de estas herramientas para saber qué marcado técnico incorporan ya de fábrica.
- **Documentar** el nivel de alfabetización en IA de tu equipo: desde febrero de 2025, garantizar un nivel suficiente de formación del personal que usa estos sistemas ya es exigible para todas las empresas, sin excepción de tamaño.

Sanciones

Los incumplimientos más graves —relacionados con los usos de IA de alto riesgo o prácticas prohibidas— pueden alcanzar hasta **35 millones de euros o el 7% de la facturación global anual**, lo que sea mayor. Las obligaciones de transparencia del artículo 50 tienen un régimen sancionador propio, generalmente menor, pero igualmente vinculante desde el 2 de agosto de 2026.

En España, la Agencia Española de Supervisión de la Inteligencia Artificial (AESIA) es la autoridad de referencia y desde agosto de 2026 dispone de plenas competencias de inspección y sanción.

Cómo protegerte: el protocolo Solutech de verificación

Frente a un panorama donde el contenido ya no se puede verificar "a simple vista", la defensa tiene que trasladarse al **proceso**, no al contenido. Este es el protocolo de cuatro capas que recomendamos implantar:

Capa 1 — Verificación fuera de banda

Ninguna instrucción de pago, cambio de cuenta bancaria o acceso urgente se ejecuta basándose solo en el canal por el que llega (email, videollamada, nota de voz). Se verifica siempre por un **canal alternativo ya conocido de antemano** (por ejemplo, llamando al número de teléfono que ya tenías guardado del proveedor, no al que aparece en el mensaje sospechoso).

Capa 2 — Doble validación para operaciones críticas

Toda transferencia por encima de un umbral definido, todo cambio de datos bancarios y todo acceso privilegiado requiere la aprobación de una segunda persona, sin excepciones aunque la petición venga "de arriba".

Capa 3 — Autenticación multifactor resistente a phishing

MFA en todos los puntos de acceso (correo, VPN, paneles de administración, banca online), priorizando métodos resistentes a phishing como FIDO2 o tokens físicos frente a los SMS, más vulnerables a interceptación.

Capa 4 — Formación continua con casos reales

La formación puntual una vez al año no funciona frente a una amenaza que evoluciona cada trimestre. Recomendamos simulacros periódicos de phishing e ingeniería social, con casos basados en incidentes reales recientes, no en ejemplos genéricos de hace cinco años.

La pregunta que detiene el 95% de los intentos

Ante cualquier comunicación que combine **urgencia + confidencialidad + una petición fuera de proceso**, una sola pregunta basta como primer filtro: *"¿Puedo confirmar esto por otro canal antes de actuar?"* Si la respuesta genera resistencia o más presión ("no hay tiempo", "no se lo digas a nadie"), es la señal de alarma más fiable que existe hoy contra el fraude con IA.

Checklist de defensa contra IA maliciosa

- [ ] Establecer un protocolo de verificación fuera de banda para cualquier instrucción de pago o cambio de cuenta bancaria.
- [ ] Implantar doble validación humana para transferencias por encima de un umbral definido.
- [ ] Activar MFA resistente a phishing (FIDO2 o token físico) en correo, VPN y paneles de administración.
- [ ] Revisar y limitar el acceso de proveedores externos a sistemas críticos, con caducidad automática.
- [ ] Configurar backups bajo el esquema 3-2-1-1-0 (tres copias, dos soportes, una externa, una inmutable, cero errores verificados con restauración real).
- [ ] Inventariar qué herramientas de IA generativa usa tu empresa y quién valida su output antes de publicarlo.
- [ ] Documentar la formación en IA de tu equipo (obligación exigible desde febrero de 2025).
- [ ] Revisar qué contenido de tu empresa (vídeos, notas de voz de directivos) está expuesto públicamente y podría usarse para clonación.
- [ ] Realizar al menos un simulacro de phishing/vishing con casos reales recientes al trimestre.
- [ ] Definir por escrito el canal alternativo de verificación para cada proveedor y directivo clave.

Preguntas frecuentes

1. ¿Cómo puedo saber si un vídeo o audio es un deepfake?
Fiarse solo del contenido ya no es fiable: la calidad de generación mejora cada mes. La defensa real está en el proceso de verificación (canal alternativo, doble validación), no en detectar fallos visuales o sonoros.

2. ¿Mi empresa pequeña puede ser objetivo de un fraude con deepfake?
Sí. Los ataques automatizados no discriminan por tamaño de empresa: buscan vulnerabilidad y urgencia operativa, no volumen de facturación.

3. ¿Qué hago si recibo una llamada o videollamada sospechosa de un directivo pidiendo una transferencia urgente?
Cuelga y verifica por el canal alternativo ya conocido (el teléfono que ya tenías guardado, no el que te ha dado la llamada). Nunca actúes solo con la información del propio mensaje sospechoso.

4. ¿El Reglamento Europeo de IA me afecta si mi empresa solo usa ChatGPT o Claude para redactar textos?
El uso auxiliar y menor de estas herramientas no suele activar la obligación de etiquetado. Si generas contenido sustancialmente sintético (imágenes, vídeos, audio) que podría confundirse con contenido real, sí debes valorar la obligación de transparencia del artículo 50.

5. ¿Qué es el fraude BEC?
Business Email Compromise: compromiso del correo corporativo para hacerse pasar por un directivo o proveedor y solicitar transferencias o cambios de datos bancarios fraudulentos.

6. ¿Pagar el rescate de un ransomware garantiza recuperar mis datos?
No. INCIBE recomienda no pagar nunca: no hay garantía de recibir la clave de descifrado, y pagar confirma a los atacantes que tu empresa es un objetivo rentable para futuros ataques.

7. ¿Qué es la doble extorsión en ransomware?
El atacante primero roba (exfiltra) los datos y después cifra los sistemas. Recuperar los archivos cifrados no elimina el riesgo de que los datos robados se publiquen o vendan igualmente.

8. ¿Necesito contratar un seguro de ciberriesgo?
Puede ser una capa adicional razonable, pero cada vez más aseguradoras exigen medidas básicas activas (MFA, backups verificados, auditorías) como condición para cubrir un incidente. El seguro no sustituye a la prevención.

9. ¿Qué hago si mi empresa es proveedora de una empresa más grande?
Asume que tu nivel de seguridad puede ser auditado como parte de la cadena de suministro de tus clientes. Reforzar tus propias defensas ya no es solo protegerte a ti: es una condición para conservar esos contratos.

10. ¿Desde cuándo es obligatoria la formación en IA de mi equipo?
Desde febrero de 2025, garantizar un nivel suficiente de alfabetización en IA del personal que use estos sistemas es exigible para todas las empresas, sin excepción de tamaño.

11. ¿Qué sanciones puedo enfrentar si no cumplo el Reglamento Europeo de IA?
Los incumplimientos más graves (usos de alto riesgo o prácticas prohibidas) pueden alcanzar hasta 35 millones de euros o el 7% de la facturación global. Las obligaciones de transparencia del artículo 50 tienen un régimen sancionador propio, generalmente inferior.

12. ¿Qué autoridad supervisa el cumplimiento del Reglamento de IA en España?
La Agencia Española de Supervisión de la Inteligencia Artificial (AESIA), con plenas competencias de inspección y sanción desde agosto de 2026.

Conclusión

La misma tecnología que en nuestra guía anterior te ayudaba a ahorrar horas de trabajo cada semana es, en manos de un atacante, la razón por la que el fraude con IA ha crecido más de un 1.200% en un año. No hace falta entender de ingeniería de modelos de lenguaje para defenderte: hace falta trasladar la verificación del contenido al proceso, formar al equipo con casos reales y no esperar a que la ley te obligue para revisar cómo usa tu empresa estas herramientas.

El mismo principio que cerraba nuestra guía de productividad con IA sigue aplicando aquí: empieza por lo pequeño y medible. En este caso, por una sola pregunta que puedes implantar hoy mismo en tu empresa: *ante cualquier petición urgente y confidencial, ¿puedo confirmar esto por otro canal antes de actuar?*