VECTRA

https://es.vectra.ai/topics/ai-scams

1-Estafas relacionadas con la inteligencia artificial

Información clave

- **Las estafas relacionadas con la inteligencia artificial aumentaron un 1210 % en 2025**, superando con creces el crecimiento del 195 % de los fraudes tradicionales, y se prevé que las pérdidas alcancen los 40 000 millones de dólares en 2027.
- Actualmente, existen **siete tipos distintos de estafas basadas en IA** que tienen como objetivo a las empresas, entre las que destacan la suplantación de identidad mediante vídeos deepfake, la clonación de voces mediante IA y el compromiso del correo electrónico empresarial (BEC) basado en IA, que suponen el mayor riesgo para las organizaciones.
- **Las defensas tradicionales están fallando.** phishing generado por IA phishing los errores gramaticales, los mensajes genéricos y las limitaciones manuales en los que se basaban los filtros de correo electrónico tradicionales y la formación en materia de concienciación para detectar el fraude.
- **La detección basada en el comportamiento llena ese vacío.** La detección y respuesta de red (NDR) y la detección y respuesta de amenazas de identidad (ITDR) detectan los patrones anómalos de red, identidad y flujo de datos que las herramientas de seguridad basadas en el contenido pasan por alto.
- **La verificación por capas ahora es obligatoria.** Los controles financieros de doble aprobación, la verificación fuera de banda y las frases de código precompartidas reducen el riesgo cuando cualquier canal de comunicación individual puede replicarse sintéticamente.

El fraude impulsado por la IA ya no es un riesgo teórico. Solo en 2024, el FBI IC3 registró pérdidas por delitos cibernéticos por valor de 16 600 millones de dólares, lo que supone un aumento interanual del 33 %, y la ingeniería social mejorada por la IA fue la causa de una parte cada vez mayor de esos incidentes. Una sola videollamada deepfake le costó a la

empresa de ingeniería Arup 25,6 millones de dólares. El phishing ahora alcanzan tasas de clics más de cuatro veces superiores a las de sus homólogos creados por humanos. Y según el informe Global Cybersecurity Outlook 2026 del Foro Económico Mundial, el 73 % de las organizaciones se vieron directamente afectadas por el fraude cibernético en 2025.

Esta guía desglosa cómo funcionan las estafas basadas en IA, los tipos más frecuentes con los que se encuentran los equipos de seguridad, los últimos datos sobre pérdidas y, lo que es más importante, cómo las empresas detectan y responden al fraude impulsado por IA cuando las defensas tradicionales no son suficientes.

¿Qué son las estafas de IA?

Las estafas con IA son esquemas fraudulentos que utilizan inteligencia artificial —incluidos grandes modelos lingüísticos, clonación de voz, generación de vídeos deepfake y agentes autónomos de IA— para engañar a las víctimas a una escala y con una sofisticación que antes eran imposibles, eliminando las limitaciones humanas que hacían que la ingeniería social tradicional fuera detectable y lenta.

Mientras que las estafas tradicionales dependían del esfuerzo, las habilidades lingüísticas y el tiempo del atacante humano, las estafas con IA eliminan por completo esas limitaciones. El atacante ya no necesita dominar el idioma del objetivo. Ya no necesita redactar manualmente mensajes personalizados. Y ya no necesita horas de preparación para un solo intento.

El Informe Internacional sobre Seguridad de la IA de 2026 reveló que las herramientas de IA que impulsan estas estafas son gratuitas, no requieren conocimientos técnicos y pueden utilizarse de forma anónima. Esa combinación —cero costes, cero habilidades, cero responsabilidad— explica por qué el fraude con IA está creciendo más rápido que cualquier otra categoría de amenazas.

Más allá de las pérdidas económicas directas, las estafas con IA crean un efecto de «degradación de la verdad». A medida que los vídeos deepfake, las voces clonadas y los textos generados por IA se vuelven indistinguibles de las comunicaciones auténticas, las organizaciones pierden la capacidad de confiar en cualquier interacción digital tal y como se presenta. Cada videollamada, mensaje de voz y correo electrónico se vuelve sospechoso.

En qué se diferencian las estafas con IA de las estafas tradicionales

El cambio fundamental es la velocidad y la calidad a gran escala. Las estafas tradicionales dependían del esfuerzo humano y contenían defectos detectables: errores ortográficos, frases incómodas, saludos genéricos. Las estafas con IA logran resultados con calidad humana a la velocidad de una máquina.

Consideremos phishing punto de referencia. Según un estudio de IBM X-Force, la IA genera un phishing convincente en cinco minutos. Un investigador humano tardaría 16 horas en redactar manualmente un correo electrónico de la misma calidad. Esto supone un aumento de velocidad de 192 veces con una calidad equivalente o superior, lo que significa que un solo atacante puede producir ahora en un día lo que antes requería un equipo de especialistas trabajando durante meses.

Las implicaciones se agravan a gran escala. La IA no solo iguala la calidad humana. Personaliza cada mensaje utilizando datos extraídos de perfiles de LinkedIn, documentos corporativos y redes sociales. Un estudio realizado en 2024 por Brightside AI reveló que phishing generados por IA lograron una tasa de clics del 54 %, en comparación con el 12 % del phishing tradicional phishing un multiplicador de eficacia de 4,5 veces.

Cómo funcionan las estafas con IA

Comprender el conjunto de herramientas del atacante es esencial para los defensores. El fraude impulsado por la IA combina múltiples tecnologías en una cadena de ataques coordinados, en la que cada etapa aprovecha diferentes capacidades de IA.

La clonación de voz representa uno de los vectores de ataque más accesibles. Una investigación de McAfee reveló que solo tres segundos de audio pueden crear un clon de voz con una precisión del 85 %. Tal y como informó Fortune en diciembre de 2025, la clonación de voz ha superado el «umbral de indistinguibilidad», lo que significa que los oyentes humanos ya no pueden distinguir de forma fiable las voces clonadas de las auténticas.

La generación de vídeos deepfake ha evolucionado desde falsificaciones evidentes hasta avatares interactivos en tiempo real. Los nuevos modelos mantienen la coherencia temporal sin el parpadeo, la distorsión o los artefactos del valle inquietante en los que se basaban los métodos de detección anteriores. El caso de Arup demostró que los participantes en vídeos deepfake pueden engañar a profesionales experimentados en llamadas en directo.

phishing basado en LLM utiliza grandes modelos lingüísticos para generar correos electrónicos hiperpersonalizados que hacen referencia a detalles específicos de la organización, transacciones recientes y estilos de comunicación individuales. Estos phishing basados en IA carecen de los signos reveladores que los filtros de correo electrónico tradicionales estaban entrenados para detectar.

Los agentes autónomos dedicados al fraude representan la última evolución. Según la investigación de Group-IB de 2026, los centros de llamadas fraudulentas basados en inteligencia artificial ahora combinan voces sintéticas, entrenamiento impulsado por LLM y respondedores de IA entrantes para ejecutar operaciones de fraude totalmente automatizadas a gran escala.

La cadena de herramientas de estafa de IA

Las cadenas de herramientas de estafa con IA ahora combinan la clonación de voz, los vídeos deepfake y los LLM oscuros en servicios comercializados que cuestan menos que una suscripción a un servicio de streaming.

El típico ataque fraudulento con IA sigue cinco etapas:

1. **Reconocimiento** : los atacantes recopilan datos públicos (redes sociales, documentos corporativos, grabaciones de conferencias) para crear perfiles de los objetivos y recopilar muestras de voz y vídeo.
2. **Generación de contenido mediante IA**: utilizando modelos de lenguaje grandes (LLM) oscuros, servicios de clonación de voz y generadores de deepfakes, los

atacantes crean phishing personalizados, mensajes de voz sintéticos o vídeos deepfake.

3. **Entrega**: el contenido generado por IA llega a los destinatarios a través del correo electrónico, llamadas telefónicas, plataformas de videoconferencia, aplicaciones de mensajería o redes sociales.
4. **Explotación**: las víctimas actúan según la comunicación fraudulenta transfiriendo fondos, compartiendo credenciales, aprobando el acceso o instalando aplicaciones maliciosas.
5. **Monetización**: los fondos robados se mueven a través de intercambios de criptomonedas, mulas de dinero o plataformas de inversión fraudulentas.

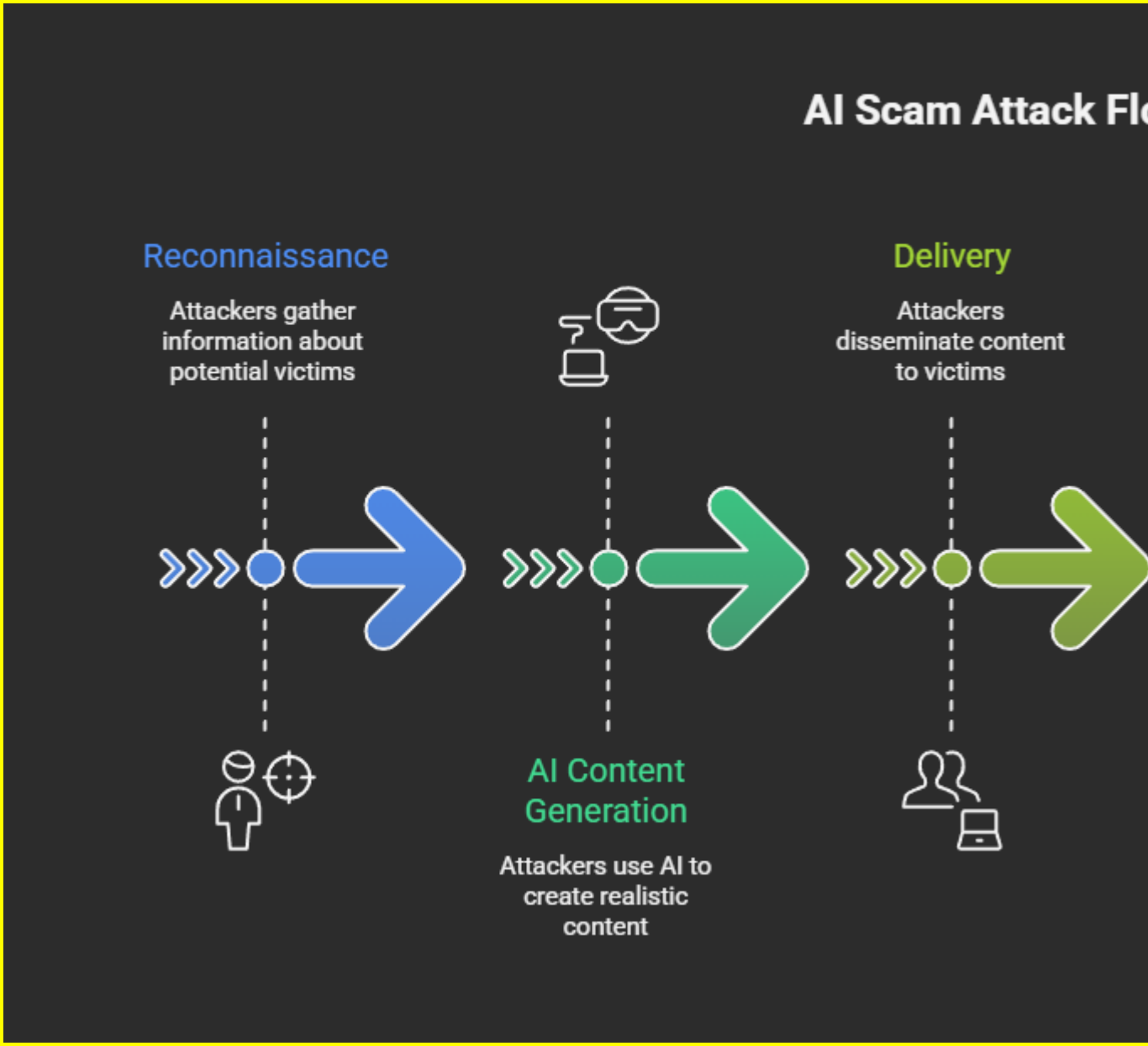

Figura: Flujo de un ataque fraudulento basado en IA. Cinco etapas secuenciales, desde el reconocimiento hasta la monetización, con la generación de contenido mediante IA como eje central. Cada nodo representa una fase concreta; las líneas muestran la progresión desde la recopilación de datos hasta la obtención de beneficios económicos.

La economía impulsa el crecimiento. Group-IB documentó kits de identidad sintética disponibles por aproximadamente 5 dólares y suscripciones a LLM oscuras que oscilan entre

30 y 200 dólares al mes. A finales de 2025, se estimaba que existían ocho millones de deepfakes en línea, frente a los aproximadamente 500 000 de 2023, lo que representa un crecimiento anual de aproximadamente el 900 %.

Las barreras de entrada han desaparecido por completo. Cualquier persona con acceso a Internet y un presupuesto reducido puede ahora lanzar campañas de ingeniería social basadas en la inteligencia artificial que hace tan solo cinco años habrían requerido recursos a nivel estatal.

Tipos de estafas relacionadas con la IA

Las estafas con IA abarcan ahora siete vectores de ataque distintos, siendo los vídeos deepfake, la clonación de voz y el BEC impulsado por IA los que plantean un mayor riesgo para las empresas. La siguiente taxonomía abarca tanto las variantes dirigidas a los consumidores como las dirigidas a las empresas.

Tabla: Clasificación de los tipos de estafas relacionadas con la IA con evaluación de riesgos empresariales. Leyenda: Tipos comunes de estafas relacionadas con la IA, sus métodos de ataque, objetivos principales, niveles de riesgo empresarial y métodos de detección recomendados.

Tipo de estafa	Método de ataque	Objetivo principal	Nivel de riesgo empresarial
Estafas con vídeos deepfake	Vídeos generados por IA que suplantan la identidad de ejecutivos en llamadas o anuncios publicitarios.	Empresas, consumidores	Crítica
Clonación de voz mediante IA (vishing)	Voz clonada utilizada en llamadas telefónicas suplantando la identidad de ejecutivos o familiares.	Empresas, consumidores	Alta
phishing generado por IA	Correos electrónicos hiperpersonalizados creados por LLM a gran escala	Empresas, consumidores	Alta
BEC impulsado por IA	BEC multimodal que combina suplantación de identidad por correo electrónico, voz y vídeo.	Empresas	Crítica

Tipo de estafa	Método de ataque	Objetivo principal	Nivel de riesgo empresarial
Fraude de identidad sintética	Identidades falsas generadas por IA que combinan datos reales y ficticios.	Servicios financieros, RR. HH.	Alta
Inversiones en IA y estafas con criptomonedas	«Expertos» generados por IA y plataformas de trading falsas	Consumidores, inversores minoristas	Medio
Estafas románticas con IA (pig butchering)	Bots con inteligencia emocional impulsados por LLM a gran escala	Consumidores	Medio

Las estafas con vídeos deepfake han aumentado un 700 % en 2025 según ScamWatch HQ, y Gen Threat Labs ha detectado 159 378 casos únicos de estafas deepfake solo en el cuarto trimestre de 2025. Las variantes empresariales incluyen la suplantación de identidad de ejecutivos en videollamadas (como en el caso de Arup), anuncios deepfake que suplantan la identidad de ejecutivos financieros y candidatos a puestos de trabajo deepfake utilizados por agentes de la RPDC.

Los ataques **de clonación de voz** mediante IA**y vishing** superan ahora las 1000 llamadas fraudulentas diarias a los principales minoristas. Más allá de dirigirse a los consumidores, los atacantes utilizan voces clonadas de ejecutivos para autorizar transferencias bancarias fraudulentas y suplantar a funcionarios del gobierno en campañas de ingeniería social.

phishing phishing spear phishing generados por IA han alcanzado un punto de inflexión. El análisis de KnowBe4 y SlashNext indica que el 82,6 % de phishing contienen ahora algún contenido generado por IA, mientras que Hoxhunt informa de que el 40 % de los correos electrónicos BEC son generados principalmente por IA. La diferencia entre estas cifras probablemente refleje la diferencia entre las metodologías de «cualquier asistencia de IA» y las de «generación totalmente por IA».

Según el FBI IC3, **los ataques de suplantación de identidad en el correo electrónico empresarial (BEC) impulsados por la inteligencia artificial** provocaron pérdidas por valor de 2770 millones de dólares en 21 442 incidentes en 2024. La inteligencia artificial está transformando los ataques BEC, que antes se limitaban al correo electrónico, en campañas multimodales que combinan correo electrónico, voz y vídeo para crear suplantaciones de identidad muy convincentes.

Las estafas relacionadas con la inversión en IA y las criptomonedas están aumentando rápidamente. La operación «Truman Show» de Check Point desplegó 90 «expertos» generados por IA en grupos de mensajería controlados, dirigiendo a las víctimas a instalar aplicaciones móviles con datos comerciales controlados por servidores. Chainalysis informó

de pérdidas por estafas con criptomonedas por valor de 14 000 millones de dólares en 2025, y las estafas basadas en IA resultaron ser 4,5 veces más rentables que el fraude tradicional.

Las estafas románticas con IA utilizan grandes modelos lingüísticos para mantener conversaciones emocionalmente inteligentes a gran escala. El informe «2026 Future of Fraud Forecast» (Previsión sobre el futuro del fraude en 2026) de Experian identifica a los bots emocionalmente inteligentes impulsados por IA como una de las principales amenazas emergentes, capaces de mantener docenas de «relaciones» simultáneas mientras adaptan el tono y la personalidad a cada objetivo.

Estafas de IA dirigidas a empresas

Las organizaciones se enfrentan a un subconjunto concentrado de tipos de estafas de IA que explotan las relaciones de confianza y los flujos de trabajo de autorización.

La suplantación de identidad de ejecutivos mediante deepfakes se centra en las transacciones de mayor valor. El incidente de Arup, en el que un empleado del departamento financiero fue engañado mediante una videollamada totalmente falsa en la que aparecía el director financiero, lo que dio lugar a 15 transacciones distintas por un total de 25,6 millones de dólares, sigue siendo el caso más destacado. Solo se descubrió gracias a una verificación manual realizada por la sede central de la empresa.

Los candidatos a puestos de trabajo falsos representan una amenaza emergente y persistente. El FBI, el Departamento de Justicia y la CISA han documentado estafas perpetradas por trabajadores informáticos de Corea del Norte que han afectado a 136 o más empresas estadounidenses, con operativos que ganan más de 300 000 dólares al año y que han llegado incluso a la extorsión de datos. Gartner prevé que uno de cada cuatro perfiles de candidatos podría ser falso en 2028.

phishing spear phishing mejorado con IA phishing gran escala se dirige a sectores industriales completos. Brightside AI documentó una campaña dirigida a 800 empresas de contabilidad con correos electrónicos generados por IA que hacían referencia a datos específicos de registro estatal, logrando una tasa de clics del 27 %, muy por encima de la media del sector para phishing .

Estafas relacionadas con la IA en cifras: estadísticas de 2024-2026

El fraude basado en la inteligencia artificial aumentó un 1210 % en 2025, y se prevé que las pérdidas alcancen los 40 000 millones de dólares en 2027, a medida que las herramientas de inteligencia artificial democratizan la ingeniería social a gran escala.

Tabla: Estadísticas sobre estafas con IA y fraudes con deepfakes, 2024-2026. Leyenda: Métricas clave sobre pérdidas financieras, volumen de ataques y prevalencia del fraude impulsado por la IA, procedentes de fuentes fidedignas.

Métrica	Valor	Fuente
Pérdidas totales por delitos cibernéticos en EE. UU. denunciadas al FBI IC3	16 600 millones de dólares (aumento interanual del 33 %)	Informe anual del FBI IC3
Pérdidas por fraude previstas relacionadas con la IA generativa	40 000 millones de dólares para 2027 (tasa compuesta de crecimiento anual del 32 % desde los 12 300 millones de dólares en 2023).	Centro Deloitte para Servicios Financieros
Aumento del fraude habilitado por IA frente al fraude tradicional	1210 % con IA frente al 195 % tradicional	Pindrop a través de la revista Infosecurity Magazine
Organizaciones afectadas por el fraude cibernético	73%	Perspectivas globales sobre ciberseguridad del Foro Económico Mundial para 2026
Pérdidas por BEC denunciadas al FBI IC3	2770 millones de dólares en 21 442 incidentes.	Informe anual del FBI IC3
Pérdidas globales por estafas (todos los tipos)	Se extrajeron 442 000 millones de dólares; el 57 % de los adultos encuestados fueron estafados.	GASA a través de ScamWatch HQ
Pérdidas por estafas con criptomonedas	14 000 millones de dólares; las estafas basadas en la inteligencia artificial son 4,5 veces más rentables.	Análisis de cadenas a través de PYMNTS
Deepfakes en línea	~8 millones (frente a ~500 000 en 2023)	DeepStrike a través de Fortune
Casos de estafas con deepfakes (solo en el cuarto trimestre de 2025)	159 378 casos únicos	Laboratorios Gen Threat
Daños causados por incidentes relacionados con deepfakes (solo en el segundo trimestre de 2025)	350 millones	Grupo IB

Métrica	Valor	Fuente
Personas que han sido víctimas de estafas con voces generadas por IA	1 de cada 4 (el 77 % de las víctimas perdió dinero)	Encuesta de McAfee a 7000 personas a través de NCOA.
Exposición al fraude en los centros de contacto	44 500 millones de dólares	Pindrop a través de la revista Infosecurity Magazine
Los líderes informan del aumento de las vulnerabilidades relacionadas con la IA	87%	Perspectivas globales sobre ciberseguridad del Foro Económico Mundial para 2026
Empresas que registran un aumento de las pérdidas por fraude (2024-2025)	Casi el 60 %	Experian a través de Fortune

Nota sobre el alcance de los datos: La cifra del FBI IC3 (16 600 millones de dólares) representa solo las denuncias presentadas ante las fuerzas del orden de EE. UU. y debe considerarse un mínimo. La cifra de GASA (442 000 millones de dólares) representa una estimación global que incluye las pérdidas no denunciadas, basada en una encuesta realizada a 46 000 adultos de 42 países. Ambas son precisas en cuanto a sus respectivas metodologías y alcance.

Estas cifras se corresponden directamente con las métricas de ciberseguridad organizativa que los CISO necesitan para informar a la junta directiva y justificar las inversiones.

Estafas relacionadas con la IA en la empresa: casos prácticos reales

Las pérdidas por estafas relacionadas con la IA empresarial van desde un fraude de deepfake de 25,6 millones de dólares en un solo incidente hasta miles de millones en pérdidas anuales por BEC, y el fraude cibernético ha superado al ransomware como la principal preocupación de los directores generales.

El informe «WEF Global Cybersecurity Outlook 2026» reveló una sorprendente desconexión en cuanto a prioridades: el fraude cibernético superó al ransomware como la principal preocupación de los directores ejecutivos en 2026, pero el ransomware sigue siendo el principal foco de atención para la mayoría de los directores de seguridad de la información. El 72 % de los líderes identificó el fraude relacionado con la inteligencia artificial como uno de los principales retos operativos, y el 87 % señaló un aumento de las vulnerabilidades relacionadas con la inteligencia artificial.

Videollamada deepfake de Arup: 25,6 millones de dólares.

En enero de 2024, un empleado del departamento financiero de la oficina de Arup en Hong Kong fue invitado a una videollamada con lo que parecía ser el director financiero de la

empresa y varios compañeros de trabajo. Todos los participantes eran deepfakes, generados a partir de imágenes de conferencias disponibles públicamente. El empleado autorizó 15 transferencias bancarias por un total de 25,6 millones de dólares (200 millones de dólares hongkoneses). El fraude solo se descubrió cuando el empleado lo verificó posteriormente con la sede central de la empresa a través de un canal independiente.

Lección aprendida: No se puede confiar únicamente en las videollamadas para la autorización financiera. Las organizaciones deben implementar controles de verificación fuera de banda y de doble aprobación para las transacciones de alto valor.

Candidatos a puestos de trabajo deepfake en Corea del Norte

El FBI ha documentado estafas perpetradas por trabajadores informáticos de Corea del Norte que han afectado a 136 o más empresas estadounidenses. Los operativos utilizan tecnología deepfake para superar entrevistas en vídeo y luego ganan más de 300 000 dólares al año, mientras desvían los ingresos hacia los programas de armamento de Corea del Norte. Algunos han pasado a la extorsión de datos, amenazando con divulgar información confidencial robada. Gartner prevé que uno de cada cuatro perfiles de candidatos podría ser falso en 2028.

Fraude de inversión «Truman Show» de Check Point

En enero de 2026, los investigadores de Check Point sacaron a la luz una operación en la que se utilizaban 90 «expertos» generados por IA para poblar grupos de mensajería controlados. Se indicaba a las víctimas que instalaran una aplicación móvil, disponible en las tiendas oficiales de aplicaciones, que mostraba datos comerciales controlados por el servidor con rendimientos falsos. Los atacantes crearon una realidad totalmente sintética para mantener el fraude.

Patrones de segmentación específicos del sector

Las diferentes industrias se enfrentan a distintos perfiles de estafas de IA. Las organizaciones de servicios financieros se enfrentan a estafas BEC, fraudes electrónicos y fraudes en centros de contacto. Un proveedor de servicios sanitarios estadounidense informó de que más del 50 % del tráfico entrante consistía en ataques impulsados por bots. Las principales empresas minoristas informan de que reciben más de 1000 llamadas fraudulentas generadas por IA al día. Las empresas de tecnología y de selección de personal informático son las más expuestas a los candidatos de empleo deepfake.

Según Cyble, el 30 % de los incidentes de suplantación de identidad corporativa de gran impacto en 2025 estuvieron relacionados con deepfakes, lo que confirma que los medios sintéticos generados por IA han pasado de ser una novedad a convertirse en un componente fundamental del fraude dirigido a las empresas. Una planificación eficaz de la respuesta a incidentes debe tener ahora en cuenta estos vectores de ataque basados en la IA.

Detección y prevención de estafas relacionadas con la IA

La defensa contra las estafas de IA empresarial requiere una detección por capas que abarque el análisis del comportamiento, la supervisión de la identidad y el análisis de la red, ya que el contenido generado por IA elude cada vez más los controles de seguridad basados en el contenido.

A continuación se presenta un marco ordenado para la defensa contra las estafas de IA empresarial:

1. **Implemente análisis de comportamiento y NDR** (detección y respuesta de red): la detección y respuesta de red identifica patrones de red anómalos asociados con la infraestructura de estafas de IA, incluidas las comunicaciones de comando y control, el tráfico de síntesis de voz y los flujos de datos inusuales.
2. **Implementar la detección de amenazas de identidad**: la detección y respuesta a amenazas de identidad (ITDR) señala patrones de autenticación anómalos, solicitudes de acceso inusuales y desviaciones de comportamiento que indican identidades comprometidas o sintéticas.
3. **Exigir controles de verificación por capas**: exigir la doble aprobación para las transacciones financieras a través de canales de comunicación independientes. Establecer frases de verificación preestablecidas para las comunicaciones de emergencia. Verificar todas las solicitudes de alto valor a través de canales fuera de banda.
4. **Mejora la formación en materia de seguridad**: cambia el enfoque de la formación, pasando de detectar errores gramaticales a reconocer la manipulación psicológica, la creación de urgencia y los contextos de solicitudes inusuales. La investigación de IBM sobre ingeniería social con IA confirma que la formación tradicional basada en «detectar errores tipográficos» ya no es eficaz.
5. **Implemente seguridad de correo electrónico mejorada con IA**: utilice filtros de correo electrónico basados en aprendizaje automático que analizan patrones de comportamiento en lugar de solo firmas de contenido. El número 9 de Microsoft Cyber Signals detalla cómo el análisis de comportamiento detecta lo que las herramientas basadas en firmas pasan por alto.
6. **Implementar la autenticación multifactorial en todas partes**: asegúrate de que la MFA cubra todos los puntos de acceso, dando prioridad a los métodos phishing(FIDO2, tokens de hardware) para las cuentas con privilegios elevados.
7. **Acepte las limitaciones de la detección de deepfakes**: la detección de deepfakes basada en el contenido es cada vez menos fiable a medida que mejora la calidad de la generación. Gartner prevé que, para 2026, el 30 % de las empresas considerarán que las soluciones de verificación de identidad independientes no son fiables por sí solas. La detección de amenazas basadas en el comportamiento proporciona una capa complementaria fundamental.

MITRE ATT&CK para amenazas de estafas relacionadas con la IA

La asignación de las técnicas de estafa con IA al MITRE ATT&CK ayuda a los equipos de GRC y a los arquitectos de seguridad a integrar los riesgos de fraude con IA en los modelos de amenazas existentes.

Tabla: MITRE ATT&CK relevantes para las estafas impulsadas por IA. Leyenda: Correspondencia entre los métodos de ataque de las estafas de IA y MITRE ATT&CK con orientación para su detección.

Táctica	Técnica ID	Nombre de la técnica	Relevancia de la estafas relacion con la IA
Acceso inicial	T1566	Phishing	phishing de spear phishing generado IA (T1566.001, T1566.002), phishir phishing servicios d mensajería (T1566. vishing mediante clonación de voz c (T1566.004)
Reconocimiento	T1598	Phishing información	Recopilación de información mejora mediante IA a trave phishing (T1598.00 enlaces (T1598.003 llamadas de voz (T1598.004).
Desarrollo de recursos	T1588.007	Obtener capacidades: Inteligencia artificial	Adquisición de mo de lenguaje grande oscuros, herramien clonación de voz, generadores de deepfakes y agente autónomos de esta

Panorama normativo para el fraude relacionado con la IA

El entorno normativo para el fraude relacionado con la IA se está endureciendo rápidamente.

- **Perfil de IA cibernética del NIST (IR 8596)**: publicado el 16 de diciembre de 2025, este borrador de marco aborda tres áreas directamente aplicables a la defensa contra estafas de IA: proteger los componentes del sistema de IA, llevar a cabo una ciberdefensa habilitada por IA y frustrar los ciberataques habilitados por IA (NIST IR 8596).
- **Medidas coercitivas de la FTC**: la FTC ha emprendido múltiples acciones contra estafas basadas en la inteligencia artificial, incluidos casos relacionados con herramientas de inversión falsas que defraudaron a los consumidores por un valor de al menos 25 millones de dólares.
- **Orden ejecutiva de la Casa Blanca sobre IA**: la orden ejecutiva de diciembre de 2025 sobre «Garantizar un marco normativo nacional para la inteligencia artificial» establece un marco regulador federal para la IA y crea un grupo de trabajo sobre litigios relacionados con la IA.

- **Leyes estatales sobre IA**: el 1 de enero de 2026 entraron en vigor varias leyes estatales, incluidas las normativas de California y Texas. La ley S.B. 24-205 de Colorado, que aborda las obligaciones en materia de gobernanza de la IA, entrará en vigor en junio de 2026.

Enfoques modernos para la defensa contra estafas relacionadas con la IA

El sector está convergiendo en un paradigma de defensa que utiliza la IA para contrarrestar la IA. Los enfoques actuales incluyen el análisis del comportamiento, la detección de amenazas a la identidad, el análisis del tráfico de red, la seguridad del correo electrónico basada en IA, las herramientas de detección de deepfakes y las plataformas avanzadas de formación en concienciación sobre seguridad.

Hay varias tendencias que están dando forma al panorama actual. La detección unificada en redes, cloud, identidades y superficies SaaS está sustituyendo a las herramientas aisladas que solo supervisan una única superficie de ataque. Los deepfakes interactivos en tiempo real plantean retos que el análisis de contenido estático no puede resolver. Y la IA agencial (sistemas de IA autónomos que actúan en nombre de los usuarios) introduce nuevos vectores de fraude en los que las máquinas manipulan a otras máquinas.

Las señales de inversión confirman la urgencia. Adaptive Security recaudó un total de 146,5 millones de dólares en financiación, incluida la primera inversión en ciberseguridad de OpenAI, centrada específicamente en la defensa contra la ingeniería social impulsada por la IA. El 94 % de los líderes encuestados por el FEM esperan que la IA sea la fuerza más importante en materia de ciberseguridad en 2026.

Fechas clave que deben tener en cuenta los defensores: la fecha límite para la declaración de política de la FTC, el 11 de marzo de 2026; el informe anual previsto del FBI IC3 2025, en abril de 2026; y la implementación de la ley S.B. 24-205 de Colorado, en junio de 2026.

Cómo Vectra AI la detección de estafas mediante IA

El enfoque Vectra AI se centra en detectar los comportamientos de red e identidad que indican que las campañas de estafa impulsadas por IA han avanzado más allá de la etapa inicial de ingeniería social. Mediante la supervisión de comunicaciones de comando y control anómalas, patrones de autenticación inusuales y flujos de exfiltración de datos asociados con la infraestructura de fraude de IA, Attack Signal Intelligence llena el vacío donde las defensas basadas en el contenido y el juicio humano fallan cada vez más frente a los ataques generados por la IA. Esto se corresponde con la filosofía de «asumir el compromiso»: encontrar a los atacantes que ya se encuentran dentro del entorno es más fiable que prevenir todos los intentos de ingeniería social mejorados por la IA en el perímetro.

Conclusión

Las estafas con IA representan la categoría de fraude de más rápido crecimiento en ciberseguridad, impulsadas por herramientas que son gratuitas, accesibles y anónimas. Los

datos son inequívocos: un crecimiento del 1210 % en el fraude habilitado por IA, 40 000 millones de dólares en pérdidas previstas para 2027 y el 73 % de las organizaciones ya afectadas.

Las empresas que se defienden con éxito contra esta amenaza comparten características comunes. Implementan una detección por capas en la red, la identidad y el correo electrónico, en lugar de confiar en un único control. Implementan flujos de trabajo financieros de doble aprobación que no confían en un único canal de comunicación. Forman a sus equipos para que reconozcan los patrones de manipulación psicológica en lugar de los errores gramaticales. Y aceptan la realidad de asumir el compromiso: la ingeniería social generada por la IA a veces tendrá éxito, por lo que la detección y la respuesta rápidas son tan importantes como la prevención.

Para los equipos de seguridad que evalúan su preparación, el marco es claro. Compare su exposición a las estafas de IA con las MITRE ATT&CK documentadas anteriormente. Evalúe si su pila de detección actual cubre las anomalías de comportamiento de la red, las amenazas de identidad y los ataques basados en el correo electrónico. Y asegúrese de que sus manuales de respuesta a incidentes tengan en cuenta phishing de deepfake, clonación de voz y phishing generado por IA.

Descubra cómo la plataformaVectra AI detecta los comportamientos de red e identidad que indican campañas de estafa impulsadas por IA, captando lo que las defensas basadas en contenido pasan por alto.

Artículos relacionados
Previsiones sobre ciberseguridad para 2026: IA, agentes y defensa del SOC
Cómo los actores maliciosos convirtieron la IA en un arma
Cómo la IA está alimentando la ciberdelincuencia y por qué crecen las brechas de seguridad
La IA es ahora la superficie de ataque: Por qué su pila de seguridad debe adaptarse rápidamente

Preguntas frecuentes

¿Cómo se puede saber si un vídeo es un deepfake?

Busque inconsistencias sutiles en la iluminación, la sincronización labial y las microexpresiones faciales, especialmente alrededor de los ojos, la línea del cabello y la mandíbula. Los errores de sincronización audiovisual y los patrones de parpadeo poco naturales también pueden indicar contenido sintético. Sin embargo, los modelos deepfake más recientes eliminan cada vez más estos artefactos visuales, lo que hace que la detección basada en el contenido no sea fiable como enfoque independiente. Los defensores de las empresas no deben confiar únicamente en la inspección visual. Las señales de comportamiento son indicadores más fiables: solicitudes inusuales, situaciones de urgencia, desencadenantes de transacciones financieras y comunicación a través de canales inesperados. En caso de duda, verifique la identidad de cualquier participante en una videollamada a través de un canal de comunicación independiente y preestablecido antes de autorizar cualquier acción. Gartner prevé que el 30 % de las empresas considerarán poco fiable la verificación de identidad independiente para 2026.

¿Qué debes hacer si eres víctima de una estafa de IA?
¿Puede la IA estafarte por teléfono?
¿Qué es el sacrificio de cerdos?
¿Cómo funcionan las estafas románticas con IA?
¿Qué es el fraude de identidad sintética?
¿Puede la IA crear sitios web falsos?

Fundamentos relacionados con la ciberseguridad

Prompt injection: tipos, vulnerabilidades CVE reales y medidas de defensa para empresas
Prompt injection el riesgo n.º 1 de los modelos de lenguaje grande (LLM) según OWASP. Descubre los tipos, los CVE reales, las estrategias de defensa y los marcos de cumplimiento para proteger tus implementaciones de IA.
Seguir leyendo

Seguridad de la IA: riesgos, marcos y buenas prácticas para 2026
Descubre qué es la seguridad de la IA, los riesgos que plantea para las empresas y las mejores prácticas para proteger los sistemas de IA. Se tratan los marcos de trabajo, las amenazas y las tendencias para 2026.
Seguir leyendo

Explicación de la gestión del estado de seguridad de la IA (AI-SPM)
Descubre qué es la gestión de la postura de seguridad de la IA (AI-SPM), cómo funciona, en qué se diferencia de la CSPM y la DSPM, y por qué es importante para proteger los sistemas de IA de las empresas.
Seguir leyendo

Explicación de la seguridad de la IA agencial: amenazas, marcos y defensas
Descubre cómo proteger los agentes de IA autónomos contra prompt injection, el uso indebido de herramientas y el abuso de identidad. Se tratan los 10 principales riesgos de OWASP, casos reales y la implementación práctica.
Seguir leyendo

phishing con IA: cómo los atacantes consiguen tasas de clics del 54 % en 5 minutos
Descubra cómo la IA está transformando phishing con tasas de clics del 54 %. Descubra métodos de detección, mapeos MITRE y estrategias de defensa para equipos de seguridad empresarial.
Seguir leyendo

MITRE ATLAS: marco de seguridad basado en IA con 16 tácticas y 84 técnicas
MITRE ATLAS (Panorama de amenazas adversarias para sistemas de inteligencia artificial) recoge 16 tácticas y 84 técnicas relacionadas con la seguridad de la IA.

Herramientas gratuitas, casos prácticos y orientación para centros de operaciones de seguridad (SOC).
Seguir leyendo

Herramientas de gobernanza de la IA: Guía de selección y seguridad para 2026

Aprenda a evaluar y seleccionar herramientas de gobernanza de IA para su empresa. Abarca el cumplimiento de la Ley de IA de la UE, la detección de IA oculta, la gobernanza de IA agencial y las mejores prácticas de implementación.
Seguir leyendo

Equipos rojos de IA: explicación de herramientas, marcos y estrategias de ataque

Descubra qué es el red teaming con IA, en qué se diferencia del red teaming tradicional, cuáles son las herramientas clave, como PyRIT y Garak, y cómo crear un programa eficaz de pruebas de seguridad con IA.
Seguir leyendo

Seguridad GenAI: cómo proteger los LLM de los ataques impulsados por IA

Descubre cómo la seguridad de GenAI protege a las empresas frente a riesgos asociados a los modelos de lenguaje grande (LLM), como prompt injection la fuga de datos. Se tratan el OWASP Top 10, el MITRE ATLAS y las estrategias de detección.
Seguir leyendo