Phishing con IA: el mensaje de tu banco que ya no tiene errores

Cuando la inteligencia artificial aprende a mentir mejor que cualquier humano

Hasta hace poco, los intentos de engaño digital tenían un defecto casi universal: se notaban. El correo del "banco" tenía faltas de ortografía. El SMS de "Hacienda" usaba un tono extraño. El mensaje del "proveedor" no conocía tu nombre real. Esos errores no eran accidentales; eran el filtro natural que separaba a quienes caían del engaño de quienes lo detectaban a tiempo.

Ese filtro ha desaparecido.

En 2026, la inteligencia artificial ha convertido el phishing —el ataque de suplantación más antiguo y extendido del mundo digital— en algo cualitativamente distinto. No es una mejora incremental. Es un salto de categoría. Y entender exactamente qué ha cambiado, y por qué ahora te afecta a ti aunque jamás hayas cometido un descuido grave, es el primer paso real hacia la protección.

Qué es el phishing y por qué sigue siendo el ataque más rentable del mundo

El phishing es la técnica de engañarte haciéndose pasar por alguien de confianza para que entregues algo valioso: una contraseña, un número de tarjeta, acceso a una cuenta, o simplemente un clic en el lugar equivocado. El nombre viene del inglés *fishing* —pescar— porque la lógica es exactamente esa: lanzar el anzuelo y esperar a que alguien pique.

Lo que lo hace el ataque más rentable del ecosistema criminal digital no es su sofisticación técnica, sino su eficiencia económica. No requiere vulnerar servidores. No exige conocimientos avanzados de programación. Solo necesita que una persona, en un

momento de distracción o confianza, haga lo que le piden. Y eso, a escala masiva, genera un retorno que ningún otro método puede igualar.

Según los datos más recientes del sector, más del 80% de los incidentes de ciberseguridad que afectan a empresas comienzan con alguna forma de phishing. En el caso de los usuarios domésticos, es el vector de entrada de la mayoría de fraudes bancarios, robos de identidad y secuestros de cuentas. Esto no es tecnología de nicho. Es el arma cotidiana del cibercrimen globalizado.

El antes y el después: cómo la IA ha transformado el ataque

Para entender la magnitud del cambio, hay que ver el proceso completo de construcción de un ataque de phishing, tanto en su versión clásica como en su versión con IA.

La versión clásica: volumen bruto, calidad baja

El phishing tradicional funcionaba como una campaña de spam masivo. Un grupo criminal compraba o robaba una base de datos de correos electrónicos, redactaba un mensaje genérico imitando a una entidad reconocida —un banco, una plataforma de pago, un servicio de streaming— y lo enviaba a millones de personas. El mensaje no sabía tu nombre. No conocía tu banco real. No adaptaba el tono. Si el 0,1% de los destinatarios caía, el ataque ya era rentable.

El resultado era fácilmente identificable para alguien mínimamente alerta: saludos genéricos ("Estimado cliente"), errores gramaticales, dominios sospechosos, imágenes mal maquetadas, urgencia artificial y poco creíble.

La versión con IA: precisión quirúrgica, escala industrial

La IA ha eliminado prácticamente todas esas señales de alerta. Y lo ha hecho en varias dimensiones simultáneas.

Primero, la personalización. Antes de que llegue el mensaje, un sistema automatizado ha recopilado información pública sobre ti: tu nombre completo, tu empresa, tu cargo, tus publicaciones en redes sociales, los servicios que mencionas, los eventos a los que

asistes, los contactos que tienes en común con otras personas. Esto es OSINT —*Open Source Intelligence*, inteligencia de fuentes abiertas—, y lo que antes requería horas de trabajo manual de un analista humano, ahora lo ejecuta un algoritmo en segundos para miles de objetivos simultáneamente.

Segundo, la generación del mensaje. Los modelos de lenguaje actuales son capaces de redactar correos electrónicos, SMS o mensajes de WhatsApp que replican con precisión el tono de comunicación de cualquier entidad. Han sido entrenados con millones de correos reales de bancos, operadoras, organismos públicos. Saben que tu banco no usa exclamaciones. Saben que Hacienda no escribe en mayúsculas. Saben cuándo una comunicación legítima incluye tu número de referencia parcial y cuándo no.

Tercero, la adaptación dinámica. Los sistemas más avanzados ya no envían un mensaje estático. Adaptan el contenido en función de la respuesta del objetivo. Si interactúas, el sistema ajusta el discurso. Si pones objeciones, las contraargumenta. Es, en esencia, un embudo de manipulación automatizado y personalizado.

El resultado es un mensaje que no tiene errores, usa tu nombre real, menciona tu banco correcto, emplea exactamente el tono que esa entidad usa contigo, y llega en un momento plausible. La probabilidad de que alguien caiga no es del 0,1%. Es órdenes de magnitud mayor.

Anatomía de un ataque real: lo que ocurre antes de que lo veas

Imagina este escenario. Es martes por la mañana. Recibes un correo de tu banco —el tuyo, no uno genérico— informando de un acceso sospechoso desde una ciudad que no reconoces. El correo incluye los últimos cuatro dígitos de tu tarjeta. Te llama por tu nombre. El diseño es idéntico al que recibes normalmente. Hay un botón para "verificar tu identidad" y asegurar la cuenta.

¿Qué ha ocurrido antes de que ese correo llegue a tu bandeja de entrada?

1. **Recopilación de datos.** En algún momento, tu dirección de correo y posiblemente otros datos básicos estuvieron expuestos en una filtración de datos —hay miles de ellas cada año, y los datos se comercializan en mercados ilegales—. El sistema sabe que eres cliente de ese banco específico porque lo mencionaste en una red

social, porque aparece en una filtración relacionada, o porque simplemente lo ha inferido estadísticamente.

2. **Enriquecimiento del perfil.** El sistema ha cruzado esa información con fuentes públicas: LinkedIn, Instagram, registros mercantiles, páginas web. Sabe tu nombre, tu empresa, posiblemente tu ciudad.
3. **Generación del contenido.** Un modelo de IA genera el correo adaptado a ti y al tono del banco. Los últimos cuatro dígitos de tu tarjeta, si están disponibles en la filtración original, los incorpora directamente. Si no, puede inventar un número parcial verosímil —o simplemente omitirlo y que el mensaje funcione igual.
4. **Envío y redirección.** El correo llega desde un dominio que imita al del banco con una variación mínima: una letra cambiada, un guión añadido, un sufijo diferente. El enlace te lleva a una página clonada del banco, alojada en un servidor efímero que desaparecerá en pocas horas.

Todo este proceso, que antes requería semanas de trabajo humano especializado para un objetivo de alto valor, ahora se ejecuta en minutos para decenas de miles de personas simultáneamente.

Las señales que quedan: cómo detectar lo que la IA no puede falsificar

Si el mensaje en sí ya no es fiable como indicador, ¿qué nos queda? La respuesta está en el contexto y en el canal, no en el contenido.

La urgencia artificial. Independientemente de cómo esté redactado, el phishing casi siempre necesita que actúes rápido. "Tu cuenta será bloqueada en 24 horas." "Confirma antes de las 18:00 o perderás el acceso." La urgencia es el mecanismo que desactiva el pensamiento crítico. Los bancos reales raramente te imponen plazos de horas para verificar tu identidad.

La solicitud de acción que llega sin que tú la hayas iniciado. Si no has pedido un cambio de contraseña, una verificación de identidad o una confirmación de pago, desconfía de cualquier mensaje que te lo solicite. Las entidades legítimas no suelen contactarte proactivamente para pedirte que "confirmes" algo que tú no has iniciado.

El dominio del remitente, no el nombre. El nombre que aparece en el correo puede ser cualquier cosa que el atacante decida. Lo que no puede falsificar fácilmente —aunque lo intenta— es el dominio desde el que se envía. Mira la dirección completa: no el nombre visible, sino la dirección entre corchetes o después del símbolo @. Un correo de `seguridad@bancosantand3r.com` no es de Banco Santander, aunque el nombre del remitente diga "Banco Santander - Seguridad".

La URL de destino antes de hacer clic. En un ordenador, sitúa el cursor sobre el enlace sin hacer clic. Verás en la barra inferior del navegador la dirección real a la que te lleva. Si hay cualquier variación respecto al dominio oficial del banco, no hagas clic.

El canal por el que llega. Tu banco tiene tus datos de contacto reales. Pero si algo te genera dudas, el camino correcto siempre es el mismo: cierra el mensaje, abre una pestaña nueva y escribe tú mismo la dirección del banco. Nunca uses el enlace que te han enviado.

La trampa de la autenticación: cuando el engaño va más allá del clic

Hay una evolución del phishing con IA que merece atención especial: el *adversary-in-the-middle* o phishing en tiempo real. En este esquema, el atacante no se limita a redirigirte a una página falsa. Actúa como intermediario entre tú y el banco real.

Tú introduces tus credenciales en la página falsa. El sistema las captura y las usa al instante para iniciar sesión en el banco real. Cuando el banco te envía el código de verificación de dos pasos a tu móvil —como medida de seguridad—, la página falsa te lo pide también, argumentando que es parte del proceso de verificación. Tú lo introduces. El atacante lo recibe y completa la autenticación en el banco real.

El resultado: incluso con la verificación en dos pasos activada, el atacante accede a tu cuenta. No porque la medida de seguridad falle, sino porque tú, creyendo estar en el entorno correcto, has proporcionado todo lo necesario.

Esto no es ciencia ficción ni una vulnerabilidad teórica. Es una técnica documentada y activa en 2026. La implicación práctica es importante: la verificación en dos pasos sigue siendo una capa de protección valiosa —y debes tenerla activada—, pero no es suficiente si llegas a una página falsa en primer lugar.

Reflexión estratégica: el problema no es tecnológico, es epistémico

Hay una dimensión del phishing con IA que suele pasarse por alto en los análisis técnicos: no es solo un problema de seguridad informática. Es un problema de epistemología aplicada —de cómo sabemos si algo es verdad o no— en el entorno digital.

Durante décadas, hemos aprendido a confiar en las señales externas de autenticidad: el aspecto de un correo, la calidad de la redacción, la coherencia del mensaje. La IA ha convertido esas señales en irrelevantes. Lo que parece auténtico ya no es necesariamente auténtico. Y lo que parece conocerte no significa que sea quien dice ser.

Esto nos obliga a un cambio de paradigma. La pregunta correcta ya no es "¿parece legítimo este mensaje?" sino "¿he verificado por un canal independiente que este mensaje

es real?". Es un cambio sutil pero fundamental. Pasa de la evaluación pasiva del contenido a la verificación activa del origen.

Para las empresas, esto tiene implicaciones directas en los protocolos internos: ninguna transferencia bancaria, ningún cambio de datos de proveedor, ningún acceso a sistemas críticos debería autorizarse únicamente sobre la base de un mensaje o correo, independientemente de su aspecto. Debe existir un segundo canal de confirmación, fuera de banda —una llamada al número que ya tienes registrado, no al que aparece en el mensaje.

Para las familias, implica hablar explícitamente de estas dinámicas con los más jóvenes y los más mayores —los dos extremos más vulnerables— y establecer protocolos sencillos para situaciones de urgencia o solicitudes de dinero.

Lo que puedes hacer hoy: tres acciones concretas

No necesitas convertirte en experto en ciberseguridad para reducir significativamente tu exposición. Estas tres acciones tienen el mayor impacto posible con el menor esfuerzo:

Verifica siempre por canal propio. Ante cualquier mensaje que te solicite acción —clic en enlace, introducir contraseña, confirmar datos, autorizar pago—, cierra el mensaje y accede al servicio directamente desde tu navegador escribiendo la dirección oficial. Si hay algún problema real con tu cuenta, lo verás al entrar normalmente.

Activa la verificación en dos pasos con aplicación de autenticación. El SMS de verificación sigue siendo mejor que nada, pero una aplicación de autenticación (como Google Authenticator o Authy) genera códigos locales que no pueden ser interceptados en tránsito. Actívala en tu correo electrónico, tu banco si lo permite, y cualquier plataforma donde tengas información sensible.

Revisa tu huella pública con criterio. Dedica veinte minutos a buscar tu nombre en Google y revisar qué información sobre ti es pública en tus redes sociales. Cuanta menos información esté disponible públicamente, menos material tienen los sistemas de IA para construir un ataque personalizado contra ti. Ajusta la privacidad de tus perfiles y reflexiona sobre qué compartes y para qué.

Conclusión: informarse es la primera línea de defensa

El phishing con IA no va a desaparecer. Al contrario, va a perfeccionarse. Pero hay algo que ningún algoritmo puede replicar: la decisión consciente de verificar antes de actuar. Esa decisión, que se toma en segundos, es lo que separa al objetivo que cae del objetivo que no cae.

La ciberseguridad en 2026 no es solo una cuestión técnica. Es una cuestión de cultura, de hábitos y de conversaciones que tenemos —o no tenemos— en nuestras familias y en nuestros equipos. Los ataques explotan la confianza porque es el recurso más abundante y menos protegido que tenemos. Protegerla, sin volverse paranoico, es una de las habilidades más valiosas del entorno digital actual.

Si este artículo te ha resultado útil, compártelo con alguien que creas que lo necesita. Una persona más informada es una persona más difícil de engañar. Y en un ecosistema donde los ataques son cada vez más automatizados y personalizados, la información sigue siendo la mejor vacuna.

Visita el blog para acceder a más recursos gratuitos sobre seguridad digital aplicada.

Isaac Ruiz Romero.