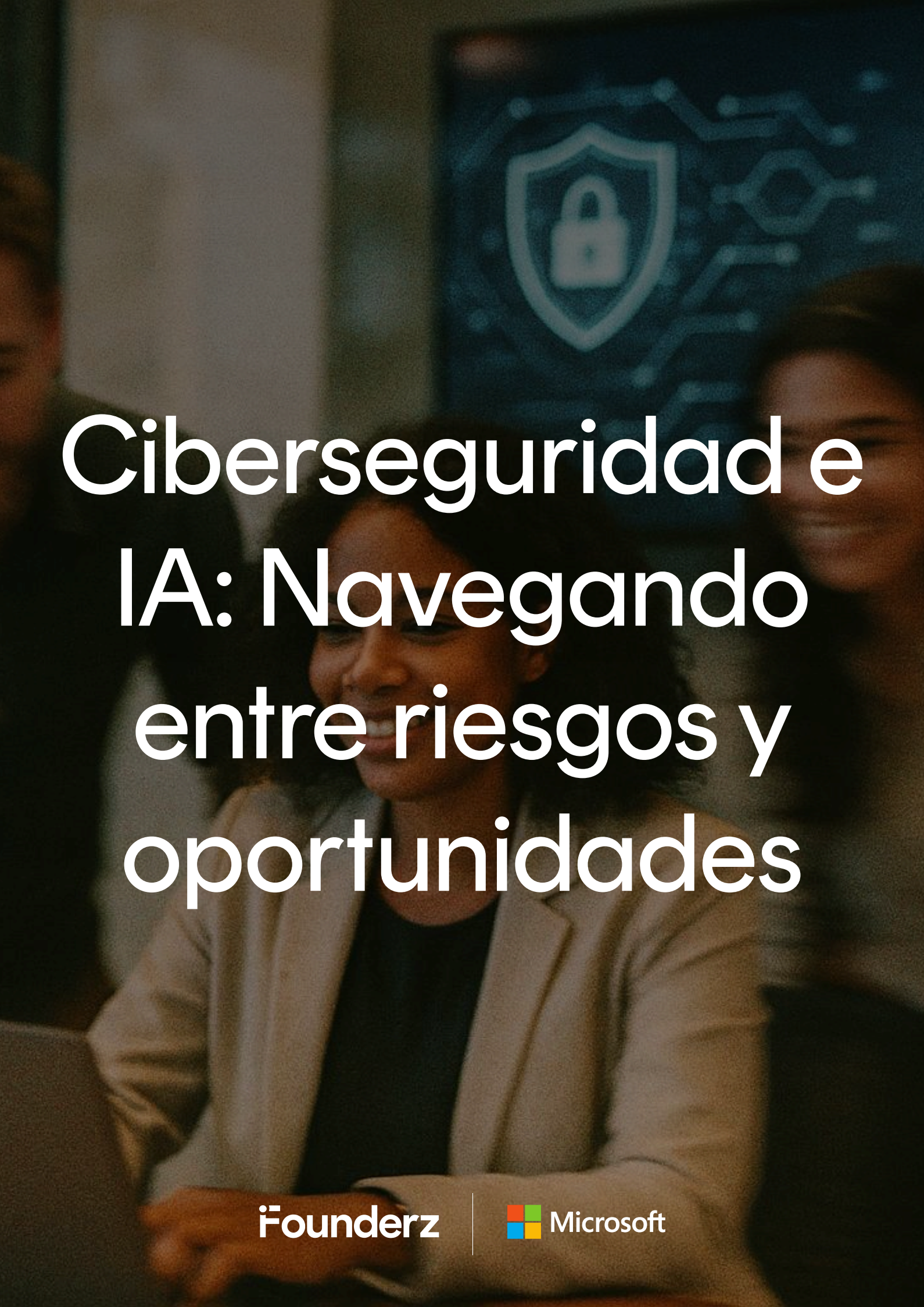

Ciberseguridad e IA: Navegando entre riesgos y oportunidades

Founderz | Microsoft

01 Deepfakes y la nueva confianza digital

Deepfakes: de la apariencia a la autenticidad

La inteligencia artificial generativa permite clonar voces e imágenes con una facilidad sin precedentes, lo que obliga a replantear cómo se entiende la confianza digital.

Clonación de identidad con pocos segundos de muestra

Hoy basta con **11 segundos de voz** o alrededor de **un minuto de vídeo** para generar un clon convincente de una persona. Ese clon puede hablar, gesticular y aparecer en situaciones que **nunca han ocurrido en la realidad.**

Estos llamados **deepfakes** hacen posible crear pruebas audiovisuales de hechos inexistentes, lo que abre la puerta a manipular opiniones, decisiones y procesos internos tanto en el ámbito personal como en el empresarial.

El cambio de paradigma en la confianza

Durante años se asumía que, si algo sonaba o parecía real, probablemente lo era. En la era de la IA, esa lógica deja de ser válida. El problema ya no es la apariencia, sino **la fiabilidad del canal** por el que se recibe la información.

- Una voz puede sonar idéntica sin pertenecer a la persona real.
- Un vídeo puede mostrar a alguien diciendo o haciendo algo que nunca sucedió.
- La intuición individual resulta insuficiente para detectar todas las falsificaciones.

Desaprender viejas creencias

El tradicional «si no lo veo, no lo creo» se transforma en «aunque se vea y se escuche, **hay que verificar**». La confianza ya no puede basarse solo en los sentidos, sino en **protocolos objetivos de validación de identidad y canal.**

El primer enemigo no es la tecnología, sino la **voz interior de exceso de confianza** que lleva a pensar que se detectará siempre a una IA. La experiencia muestra que, ante un ataque bien diseñado, cualquier persona puede ser engañada.

Impactos principales de los deepfakes en las organizaciones

La falsificación de identidad con IA genera riesgos concretos que afectan directamente a procesos financieros, reputacionales y de toma de decisiones.

Suplantación de directivos

Uso de voz o vídeo clonados de alta dirección para ordenar acciones urgentes, aprobar pagos o modificar condiciones críticas sin seguir los procesos establecidos.

Fraude financiero y desvío de pagos

Instrucciones falsas hacia equipos financieros para cambiar cuentas bancarias, validar transferencias o dar de alta proveedores inexistentes, aprovechando la confianza en una identidad aparente.

Manipulación de la reputación

Creación de vídeos o audios fabricados que muestran mensajes ofensivos, decisiones polémicas o comportamientos inaceptables atribuidos a personas clave, dañando su imagen pública.

Desinformación interna

Mensajes falsos que aparentan proceder de equipos de liderazgo y que pueden alterar decisiones, crear confusión en plantillas o romper la coordinación entre departamentos.

Erosión de la confianza general

Cuando las falsificaciones se vuelven creíbles, se debilita la confianza entre personas, equipos y socios, ralentizando operaciones y dificultando la colaboración efectiva.

De la confianza intuitiva a la verificación sistemática

La presencia de deepfakes obliga a sustituir la confianza basada en la apariencia por criterios objetivos centrados en el canal y el proceso.

Aspecto	Enfoque tradicional	Enfoque en la era de la IA
Identidad de la persona	Se asume auténtica si la voz y la imagen parecen reales.	Se valida mediante **protocolos formales** (procedimientos, claves, canales verificados).
Canal de comunicación	Importa sobre todo el contenido del mensaje.	Se prioriza si el canal está **corporativamente verificado** y controlado.
Validez de las instrucciones	Se acepta una instrucción si parece coherente y urgente.	Se exige **contraste adicional** en función del tipo de operación y su sensibilidad.
Urgencia del mensaje	La urgencia se interpreta como prueba de importancia.	La urgencia se entiende como **posible técnica de presión** propia de un intento de fraude.
Pruebas audiovisuales	Vídeos y audios se consideran evidencias casi incuestionables.	Se analizan como contenido que **puede haber sido generado o alterado** mediante IA.

Confiar únicamente en la intuición ante voces o vídeos realistas resulta insuficiente. La seguridad comienza al asumir que **cualquier persona puede ser engañada** y que la protección depende de procesos robustos.

02 Procedencia y trazabilidad en contenidos digitales

Procedencia: la historia verificable de un archivo

La procedencia permite conocer quién creó un contenido digital, cómo se ha modificado y si intervino inteligencia artificial en el proceso.

Qué son las credenciales de contenido

Las **Content Credentials** son metadatos verificables que se integran en imágenes, audio, vídeo o documentos PDF. Su objetivo es aportar una **capa de transparencia y trazabilidad** sobre el origen y la edición del archivo.

Basadas en el estándar abierto **C2PA**, estas credenciales permiten que un tercero inspeccione el contenido, vea con qué herramientas se generó y qué modificaciones se han aplicado a lo largo del tiempo.

Por qué la procedencia es relevante

Cuando un archivo incluye credenciales de procedencia, se abre la posibilidad de **auditar su historia** en lugar de juzgarlo solo por su apariencia. Esto ayuda a:

- Detectar si ha intervenido una herramienta de IA generativa en su creación.
- Identificar ediciones sustanciales que puedan alterar el significado del contenido.
- Diferenciar material original de versiones manipuladas o recortadas.

Una herramienta útil, pero parcial

La procedencia dota a los contenidos digitales de un «historial clínico», pero **no se aplica a todo lo que circula.** Muchos archivos se comparten sin credenciales, y otros pierden esta información al ser reenviados o convertidos de formato.

Por ello, la procedencia debe verse como **una capa más** dentro de una estrategia de seguridad amplia, nunca como la única línea de defensa frente a la desinformación o el fraude.

Uso básico de la procedencia en archivos pregenerados

En contenidos ya existentes, las credenciales de procedencia ayudan a evaluar si un archivo merece confianza antes de integrarlo en procesos críticos.

1. **Comprobar si el archivo incluye credenciales**: En primer lugar se verifica si la imagen, el audio, el vídeo o el PDF contienen información de procedencia basada en Content Credentials o C2PA.
2. **Revisar el historial de creación y edición**: Una vez detectadas las credenciales, se analiza quién creó el archivo, qué herramientas se usaron y qué modificaciones relevantes se han registrado.
3. **Contrastar el historial con el contexto**: Se evalúa si la historia técnica del archivo encaja con lo que se afirma sobre él (fecha, origen, autoría, propósito) o si aparecen incoherencias.
4. **Documentar la evaluación**: Para usos sensibles, resulta recomendable dejar constancia de cómo se ha verificado la procedencia y qué conclusiones se han obtenido, integrándolo en las políticas internas de seguridad.

Limitaciones prácticas de la procedencia

Aunque poderosa, la procedencia no puede cubrir todos los escenarios de riesgo, especialmente en situaciones dinámicas o maliciosas.

Ausencia de credenciales en muchos contenidos

Gran parte de los archivos que circulan por mensajería o redes sociales se generan y comparten sin incluir ninguna marca de procedencia verificable.

Pérdida de información al reenviar

La descarga, captura de pantalla o conversión de formatos puede eliminar o alterar las credenciales originales, dificultando la reconstrucción del historial del archivo.

Imposibilidad en comunicaciones en tiempo real

Llamadas telefónicas, videoconferencias o chats en directo no pueden llevar un historial completo de procedencia, ya que **están ocurriendo en el momento**.

Falsificaciones sin trazas visibles

Un atacante puede generar contenido sintético y distribuirlo sin añadir ninguna credencial, de modo que no exista una pista directa sobre su origen real.

Necesidad de enfoques complementarios

Estas limitaciones hacen necesario combinar la procedencia con **procesos internos sólidos, verificación por canales alternativos** y análisis técnico del contenido.

La procedencia ofrece una valiosa pista sobre el origen de un archivo, pero **no sustituye** a la verificación de canal, contexto y proceso. La seguridad eficaz se construye por capas.

03 Defensa por capas frente a fraudes con IA

Modelo de protección en cuatro capas

Ante la facilidad para clonar voces y vídeos, la defensa más eficaz consiste en combinar verificaciones sucesivas que dificulten enormemente el trabajo del atacante.

Capa 1: origen y canal corporativo

El primer filtro consiste en comprobar si la solicitud llega por un **canal corporativo verificado**: cuentas oficiales, herramientas internas o vías previamente registradas. Un mensaje inesperado desde un canal poco habitual merece especial cautela.

Esta capa no evita todos los ataques, pero reduce significativamente el riesgo de aceptar instrucciones provenientes de correos personales, números desconocidos o perfiles no autorizados.

Capa 2: análisis del contexto

El segundo nivel se centra en el **sentido del mensaje respecto a procesos, calendario y responsables.** Se examinan elementos como importes, plazos y tipo de operación:

- ¿Encaja la solicitud con el procedimiento habitual?
- ¿La persona que escribe suele intervenir en este tipo de temas?
- ¿El importe es sospechosamente alto... o extrañamente bajo?

Cuando el contexto no coincide con el funcionamiento normal de la organización, aumenta la probabilidad de estar ante un intento de manipulación.

Capa 3: comprobaciones técnicas del contenido

En esta fase se aplican **métodos técnicos** para evaluar el material recibido: búsqueda inversa de imágenes, análisis de sombras y reflejos en un vídeo, sincronía labial o patrones de audio.

También se observan comportamientos anómalos en comunicaciones en tiempo real. Por ejemplo, si una persona que siempre activa su cámara insiste en desactivarla alegando mala conexión en un lugar donde se sabe que hay fibra óptica en buen estado, esa incoherencia constituye una señal de alerta.

Capa 4: confirmación por un canal distinto

La capa más importante consiste en **pedir confirmación por una vía diferente** a la utilizada para enviar la solicitud inicial. Así, un correo se valida por teléfono a un número ya registrado; una llamada, mediante un mensaje en el chat interno o un correo electrónico.

El principio es sencillo: **nunca validar por el mismo canal** la instrucción que se ha recibido, especialmente cuando afecta a pagos, cambios bancarios, accesos privilegiados o datos personales.

Capas de protección y su función específica

Cada capa añade fricción para el atacante, pero mantiene un esfuerzo razonable para quienes actúan de forma legítima dentro de la organización.

Capa	Objetivo principal	Ejemplo de aplicación
1. Origen y canal	Garantizar que la solicitud llegue por una vía previamente controlada.	Aceptar instrucciones financieras solo desde cuentas de correo corporativas verificadas, nunca desde direcciones personales.
2. Contexto	Comprobar que la operación tiene sentido dentro de los procesos y calendarios establecidos.	Rechazar una transferencia urgente a un nuevo proveedor si nadie del área responsable reconoce la relación comercial.
3. Análisis técnico	Identificar posibles manipulaciones o contenidos sintéticos.	Detectar que una imagen procede de un banco de imágenes tras realizar una búsqueda inversa en Internet.
4. Canal alternativo	Confirmar que la persona real respalda la solicitud.	Llamar a un número ya guardado en la agenda corporativa para verificar un cambio de cuenta bancaria recibido por correo.
5. Código rodante	Aumentar exponencialmente la dificultad de suplantar a alguien.	Exigir una frase clave que cambia periódicamente y que solo conocen las personas autorizadas antes de aprobar operaciones sensibles.

Aplicación de la regla de dos canales y del código rodante

Las solicitudes sensibles requieren un protocolo específico que combine confirmación por vías distintas y una clave dinámica conocida solo por personas autorizadas.

1. **Identificar qué operaciones son sensibles**: Se definen como críticas las instrucciones relativas a pagos, cambios de cuenta bancaria, altas de proveedores, concesión de accesos privilegiados o intercambio de datos personales.

2. **Exigir siempre dos canales independientes**: Cualquier solicitud de este tipo debe confirmarse desde un canal distinto al original (teléfono, correo corporativo, chat interno), utilizando datos de contacto ya registrados.

3. **Incorporar un código rodante**: Se utiliza una frase o código que cambia con una cadencia fija y que solo conocen las personas autorizadas. Sin ese código vigente, la aprobación se considera no válida.

4. **Rechazar la urgencia como motivo para saltarse el proceso**: La presión del «es para ahora mismo» o «viene de la máxima dirección» se interpreta como un posible indicador de ataque, no como un permiso para obviar el protocolo.

5. **Normalizar el hábito en toda la organización**: Se comunica de forma sencilla que **ninguna aprobación** de operaciones sensibles será válida sin confirmación por segundo canal y, cuando aplique, sin el código rodante correspondiente.

El objetivo no es «cazar» atacantes, sino **evitar pérdidas y proteger el tiempo del negocio**. Si cumplir el protocolo añade unos minutos para la organización, pero supone semanas o meses de trabajo extra para un atacante, la balanza de esfuerzo se inclina claramente a favor de la defensa. Los atacantes suelen escoger la empresa con procesos más débiles; reforzar las capas de seguridad reduce la probabilidad de convertirse en el objetivo más fácil. A menudo basta con que el atacante cambie una sola pieza del «puzle» del proceso para pasar desapercibido, de ahí la importancia de revisar cada paso crítico.

Creado por Victoria, AI Founderz Fellow, y aprobado por el equipo de Founderz.

Última actualización 5 de diciembre de 2025

Este documento fue originalmente generado por la IA y revisado por nuestro equipo humano. En Founderz, utilizamos la IA de forma responsable y transparente.