Deepfakes financieros y fraude corporativo: nuevos desafíos para la auditoría y el control interno

Por: Mg. CPC. Giannina Marleni Castillo Castillo (Perú)

Mg. CPC. Juan Randall Fabián Ramírez Huerta (Perú)

Miembros de la Comisión Técnica de Investigación Contable de la Asociación Interamericana de Contabilidad – AIC

Mg. CPC. Giannina Marleni Castillo Castillo (Perú)

Mg. CPC. Juan Randall Fabian Ramírez Huerta (Perú)

Resumen

La inteligencia artificial generativa permite recrear voces, rostros, videollamadas, documentos y mensajes con un grado de realismo capaz de confundir incluso a usuarios experimentados. En el ámbito empresarial, estos contenidos sintéticos pueden emplearse para imitar a directivos, alterar instrucciones de pago, modificar cuentas bancarias, fabricar soportes documentales o influir en investigaciones internas. El presente artículo examina los *deepfakes* financieros como una amenaza emergente para la auditoría y el control interno. Se argumenta que las organizaciones ya no pueden considerar la voz, el rostro, la firma digitalizada o el estilo comunicativo de un superior como pruebas suficientes de identidad. Frente a este escenario, se requiere fortalecer la segregación de funciones, la validación por canales independientes, la autenticación reforzada y los protocolos aplicables a desembolsos excepcionales. Asimismo, la auditoría debe extender su escepticismo profesional hacia el origen, la integridad y la trazabilidad de los archivos electrónicos. Se concluye que la protección más efectiva no consiste únicamente en detectar una falsificación, sino en impedir que una comunicación aparentemente legítima pueda, por sí sola, generar una pérdida económica o alterar la información corporativa.

Palabras clave: *Deepfakes*, fraude corporativo, suplantación digital, auditoría, control interno, evidencia electrónica, inteligencia artificial.

1. Introducción

La inteligencia artificial generativa ha modificado una premisa básica de la comunicación contemporánea: ver o escuchar a una persona ya no garantiza que esa persona haya participado realmente en la interacción. Las herramientas disponibles pueden recrear gestos, movimientos, timbres de voz y formas de expresión con un nivel creciente de verosimilitud. Schmitt y Flechais (2024) explican que esta tecnología potencia la ingeniería social mediante contenido realista, personalización avanzada y automatización de ataques, mientras que Gong y Li (2024) señalan que la continua evolución de los métodos de falsificación dificulta distinguir las piezas genuinas de las manipuladas. La ruptura entre apariencia y autenticidad obliga a reconsiderar cómo se construye la confianza en los entornos digitales, pues aquello que resulta convincente para los sentidos puede carecer de sustento real.

Este fenómeno adquiere especial relevancia en las organizaciones, donde numerosas operaciones continúan dependiendo de señales personales de autoridad. Una llamada del gerente general, una reunión virtual con un ejecutivo o un mensaje redactado con su estilo habitual pueden ser interpretados como instrucciones legítimas, aun cuando no hayan sido emitidos por él. Vecchietti et al. (2025) identifican que los medios sintéticos amenazan la integridad de los datos, la continuidad de las operaciones y la privacidad empresarial, mientras que Sandoval et al. (2024) vinculan estas falsificaciones con delitos de identidad, fraude y manipulación informativa. En consecuencia, el punto débil no se limita a la infraestructura tecnológica: también reside en una cultura corporativa que suele equiparar familiaridad con veracidad.

En el campo financiero, la novedad no está en la intención de engañar, sino en la posibilidad de reproducir de forma creíble las señales utilizadas para reconocer a una autoridad. Schmitt y Flechais (2024) sostienen que la inteligencia artificial facilita

campañas de ingeniería social más precisas y adaptadas al contexto de cada víctima, mientras que Babaei et al. (2025) advierten que la generación de contenido sintético avanza con mayor rapidez que algunos mecanismos diseñados para identificarlo. Así, el delincuente puede estudiar la estructura jerárquica, imitar el lenguaje de un ejecutivo y elegir un momento de presión, como el cierre contable o una negociación confidencial, para solicitar una transferencia. El riesgo se intensifica cuando la obediencia jerárquica pesa más que la obligación de verificar.

La problemática también alcanza a la auditoría porque una grabación, una fotografía o una conversación virtual pueden presentarse como respaldo de una transacción o como prueba en una investigación. Sandoval et al. (2024) concluyen que los *deepfakes* dificultan la atribución de responsabilidades y debilitan la credibilidad de la evidencia, mientras que Gong y Li (2024) muestran que el desempeño de los detectores varía según el tipo de manipulación y los datos con los que fueron entrenados. Por ello, la ausencia de una alerta técnica no demuestra que un archivo sea auténtico. La labor profesional debe desplazarse desde la simple observación del contenido hacia la comprobación de su procedencia y del recorrido que siguió hasta llegar al auditor.

El desarrollo de contenidos sintéticos altera primero la confianza en las comunicaciones digitales; posteriormente debilita los mecanismos tradicionales de autorización y, finalmente, incrementa el riesgo de pagos fraudulentos, modificaciones no aprobadas y conclusiones sustentadas en soportes falsos. En respuesta a esta cadena de riesgos, el presente artículo analiza seis dimensiones: la suplantación de directivos, el fraude en autorizaciones, la manipulación de evidencia, la corroboración por vías independientes, la segregación de funciones y los protocolos destinados a operaciones extraordinarias.

2.1. La suplantación de directivos y la fragilidad de la autoridad digital

La posibilidad de imitar a un ejecutivo convierte su identidad digital en un activo que necesita controles específicos. Schmitt y Flechais (2024) sostienen que la inteligencia artificial puede reproducir mensajes, voces e imágenes adaptadas al perfil de cada destinatario, mientras que Vecchietti et al. (2025) advierten que la falsificación de identidades afecta la integridad informativa y la continuidad del negocio. Un atacante no necesita ingresar directamente al sistema financiero cuando puede inducir a un trabajador autorizado a ejecutar la acción por él. Por esta razón, la identidad de un directivo no debería descansar en rasgos imitables, sino en credenciales, registros y mecanismos de confirmación que puedan auditarse posteriormente.

La urgencia y la confidencialidad funcionan como instrumentos para desactivar el razonamiento crítico del receptor. Schmitt y Flechais (2024) explican que la ingeniería social asistida por inteligencia artificial puede adaptar su discurso a las emociones y respuestas de la víctima, mientras que Sandoval et al. (2024) reconocen la manipulación y el fraude como aplicaciones delictivas recurrentes de los contenidos sintéticos. Una instrucción que exige actuar de inmediato, mantener silencio o evitar los canales habituales debe elevar el nivel de cautela. En términos críticos, una organización se vuelve vulnerable cuando cuestionar a un superior se interpreta como desobediencia, incluso si la consulta busca proteger sus recursos.

La exposición pública de los ejecutivos facilita la construcción de imitaciones convincentes. Vecchietti et al. (2025) destacan la relación entre privacidad, disponibilidad de datos y riesgo empresarial, mientras que Schmitt y Flechais (2024) señalan que la personalización de los ataques se alimenta de información obtenida en entornos digitales. Entrevistas, conferencias, fotografías y publicaciones institucionales pueden proporcionar muestras suficientes para reproducir el aspecto o la voz de una autoridad. Esto no implica ocultar toda presencia corporativa, sino reconocer que la comunicación pública también crea una superficie de riesgo que debe ser administrada.

2.2. Autorizaciones fraudulentas y operaciones financieras excepcionales

La voz o la imagen de una autoridad ya no constituyen una aprobación confiable por sí mismas. Gong y Li (2024) describen avances que permiten reproducir rasgos faciales y movimientos con gran realismo, mientras que Babaei et al. (2025) muestran que los métodos de generación y reconocimiento evolucionan de manera paralela. Una videollamada puede servir como medio de comunicación, pero no debería sustituir la autorización registrada dentro del flujo institucional. El control debe demostrar quién aprobó, cuándo lo hizo, qué datos revisó y mediante qué credencial, en lugar de depender de la impresión subjetiva de un trabajador.

Los pagos fuera de la rutina concentran las condiciones más favorables para el engaño. Vecchietti et al. (2025) relacionan los *deepfakes* con amenazas para la continuidad operativa y la toma de decisiones, mientras que Schmitt y Flechais (2024) indican que la personalización permite crear relatos coherentes con situaciones reales de la empresa. Una transferencia internacional inesperada, el cambio repentino de una cuenta bancaria o una solicitud confidencial vinculada con una adquisición pueden parecer plausibles porque se apoyan en circunstancias existentes. Precisamente por su carácter excepcional, estas operaciones deberían atravesar controles más rigurosos y no procedimientos abreviados.

La modificación de los datos de un beneficiario merece un tratamiento distinto al de una transacción ordinaria. Sandoval et al. (2024) identifican la suplantación y el fraude como consecuencias relevantes del contenido manipulado, mientras que Vecchietti et al. (2025) resaltan la necesidad de proteger la integridad de los datos empresariales. Un pago correctamente autorizado puede terminar en manos equivocadas si previamente se alteró la cuenta receptora. La revisión no debe concentrarse únicamente en el importe o en la existencia de una factura; también debe comprobar que los datos maestros no fueron modificados mediante una comunicación falsa.

2.3. Evidencia audiovisual y documentos de apariencia legítima

El realismo de una pieza audiovisual ha perdido valor como criterio de autenticidad. Gong y Li (2024) sostienen que la sofisticación de las falsificaciones dificulta la diferenciación humana y automatizada, mientras que Sandoval et al. (2024) advierten que este fenómeno puede introducir pruebas fabricadas en investigaciones y procesos legales. En consecuencia, una grabación debe examinarse junto con sus metadatos, fuente original, fecha, dispositivo de creación y cadena de custodia. La apariencia puede orientar una revisión inicial, pero no debería sustentar una conclusión profesional definitiva.

Los documentos sintéticos pueden ser formalmente impecables y económicamente falsos. Schmitt y Flechais (2024) señalan que la inteligencia artificial facilita la creación de mensajes, identidades y soportes visuales convincentes, mientras que Vecchietti et al. (2025) advierten que esta capacidad compromete la integridad de la información utilizada por las organizaciones. Una factura puede incluir logotipos, firmas, sellos y datos tributarios correctos, pero corresponder a una operación inexistente. Por ello, revisar únicamente su formato es insuficiente; debe contrastarse con órdenes de compra, evidencia de recepción, registros independientes y confirmaciones externas.

La incertidumbre sobre los contenidos digitales también permite negar hechos auténticos. Sandoval et al. (2024) explican que la difusión de los *deepfakes* erosiona la confianza general en la evidencia, mientras que Altuncu et al. (2024) señalan que el objetivo de estas técnicas consiste precisamente en dificultar la distinción entre lo verdadero y lo fabricado. Una persona podría desconocer una comunicación legítima alegando que fue manipulada, incluso cuando no lo fue. Esta posibilidad exige preservar desde el origen los archivos originales, las marcas de tiempo y los registros de acceso, pues la organización debe estar preparada para probar tanto la falsedad como la autenticidad.

La detección automatizada debe utilizarse como indicio y no como veredicto. Babaei et al. (2025) destacan las limitaciones que presentan algunos detectores frente a técnicas no incluidas en su entrenamiento, mientras que Gong y Li (2024) señalan dificultades de generalización entre diferentes conjuntos de datos. Una herramienta puede producir falsos positivos o no reconocer una modalidad reciente de manipulación. El juicio profesional exige integrar sus resultados con otros elementos y documentar por qué se acepta o rechaza la autenticidad del material examinado.

2.4. Corroboración independiente y ruptura de la narrativa fraudulenta

La confirmación por una vía separada reduce el poder persuasivo de una comunicación falsificada. Schmitt y Flechais (2024) recomiendan combinar defensas tecnológicas con procedimientos orientados al comportamiento humano, mientras que Vecchietti et al. (2025) enfatizan la importancia de la gobernanza y la integridad de los datos frente a los riesgos sintéticos. Una solicitud sensible debe comprobarse mediante un número, contacto o plataforma previamente registrada, nunca utilizando los datos incluidos en el propio mensaje sospechoso. La independencia del medio es esencial porque obliga a validar la orden fuera del entorno controlado por el atacante.

La corroboración debe diseñarse antes del incidente para evitar decisiones improvisadas bajo presión. Sandoval et al. (2024) plantean la necesidad de fortalecer las respuestas institucionales ante la suplantación, mientras que Schmitt y Flechais (2024) muestran que los atacantes aprovechan la urgencia y el contexto para influir en la conducta. Si el trabajador debe decidir en el momento a quién llamar o qué pasos seguir, puede terminar consultando al mismo canal comprometido. Los contactos autorizados, preguntas de verificación y rutas de escalamiento deben estar definidos y actualizados previamente.

La confirmación independiente pierde eficacia cuando se convierte en una formalidad predecible. Gong y Li (2024) advierten que las técnicas de falsificación

continúan adaptándose a los mecanismos de defensa, mientras que Vecchietti et al. (2025) señalan que la gestión del riesgo requiere respuestas organizacionales dinámicas. Un atacante que conoce el procedimiento puede preparar respuestas para controles rutinarios o imitar a varios participantes. La validación debe combinar factores distintos, credenciales, registros, contacto alternativo y revisión contextual, en lugar de depender siempre de una misma pregunta o llamada.

2.5. Segregación de funciones como barrera frente al engaño

La separación de responsabilidades limita el daño cuando uno de los participantes ha sido manipulado. Schmitt y Flechais (2024) resaltan que las personas siguen siendo un objetivo central de la ingeniería social, mientras que Vecchietti et al. (2025) consideran que la gestión del riesgo de *deepfakes* debe proteger la continuidad y la integridad operativa. Si una sola persona puede crear un proveedor, modificar su cuenta, registrar la obligación y liberar el desembolso, basta con convencerla para completar el fraude. Un sistema bien diseñado no supone que nadie será engañado; procura que el error de un individuo no sea suficiente para producir el perjuicio.

La doble aprobación solo es útil cuando los revisores actúan con verdadera independencia. Sandoval et al. (2024) muestran que las falsificaciones pueden emplearse para manipular a distintas personas, mientras que Schmitt y Flechais (2024) describen la capacidad de automatizar campañas personalizadas a gran escala. Dos firmas no ofrecen protección real si el segundo aprobador se limita a confirmar que la primera ya existe. La revisión adicional debe incluir el análisis del beneficiario, la justificación económica, los antecedentes y las señales de anomalía.

Los accesos privilegiados requieren controles compensatorios más exigentes. Vecchietti et al. (2025) subrayan que la integridad de los datos constituye una defensa relevante frente al fraude sintético, mientras que Gong y Li (2024) evidencian la rapidez con que evolucionan estas amenazas. Quienes administran perfiles, maestros de proveedores o parámetros financieros pueden eludir controles ordinarios debido a la naturaleza de sus funciones. En estos casos, la organización necesita bitácoras inalterables, alertas automáticas, revisiones posteriores y restricciones temporales que reduzcan la posibilidad de abuso o manipulación.

2.6. Protocolos para pagos extraordinarios

Toda operación atípica debe activar una pausa deliberada antes de su ejecución. Schmitt y Flechais (2024) explican que la urgencia es utilizada para limitar el análisis y acelerar respuestas, mientras que Sandoval et al. (2024) advierten que la suplantación puede conducir a decisiones fraudulentas con consecuencias difíciles de revertir. Un periodo mínimo de espera permite corroborar la solicitud, revisar antecedentes y consultar a otros responsables. La pausa no constituye una demora innecesaria, sino una barrera frente a decisiones tomadas bajo presión artificial.

Los umbrales monetarios deben complementarse con criterios cualitativos de riesgo. Vecchietti et al. (2025) destacan que los efectos de los *deepfakes* trascienden la privacidad y alcanzan la continuidad empresarial, mientras que Schmitt y Flechais (2024) muestran que los ataques pueden adaptarse al contexto y a los límites conocidos. Dividir una suma elevada en operaciones menores podría evitar un control basado exclusivamente en el

importe. Por ello, también deben considerarse cambios de beneficiario, horarios inusuales, jurisdicciones nuevas, instrucciones secretas y desviaciones frente al comportamiento histórico.

La organización debe disponer de una ruta inmediata para detener o recuperar desembolsos sospechosos. Sandoval et al. (2024) resaltan que las respuestas institucionales son esenciales frente a la falsificación y el fraude, mientras que Vecchietti et al. (2025) relacionan la preparación organizacional con la continuidad de las operaciones. La reacción no puede comenzar con la búsqueda improvisada de responsables después de descubrir la pérdida. Deben estar definidos los contactos bancarios, las facultades para suspender pagos, la preservación de evidencias y las responsabilidades de comunicación.

Tabla 1

Modalidades de fraude sintético y controles preventivos

Modalidad	Situación corporativa	Debilidad aprovechada	Respuesta recomendada
Clonación de voz	Llamada atribuida a un ejecutivo	Confianza en el reconocimiento auditivo	Devolver la llamada a un número registrado
Videollamada sintética	Reunión virtual con supuestos directivos	Confianza en la apariencia	Confirmación adicional en el sistema institucional
Mensaje personalizado	Orden redactada con el estilo de un superior	Jerarquía, urgencia o reserva	Verificación mediante una vía diferente
Cambio de cuenta	Solicitud falsa atribuida a un proveedor	Falta de control sobre datos maestros	Confirmación con un contacto autorizado
Documento fabricado	Factura o contrato de aspecto legítimo	Revisión limitada al formato	Contraste con la realidad de la operación
Grabación alterada	Archivo usado en una investigación	Presunción de veracidad audiovisual	Análisis forense y comprobación de procedencia
Desembolso excepcional	Transferencia urgente y confidencial	Omisión temporal de procedimientos	Pausa obligatoria y doble evaluación independiente

Fuente: elaboración propia con base en Schmitt y Flechais (2024), Gong y Li (2024), Sandoval et al. (2024) y Vecchietti et al. (2025).

2.7. Implicaciones para la auditoría interna y externa

El escepticismo profesional debe abarcar tanto el contenido de la evidencia como su procedencia. Sandoval et al. (2024) sostienen que los *deepfakes* amenazan la integridad de los soportes y la atribución de responsabilidades, mientras que Gong y Li (2024) evidencian las dificultades técnicas para reconocer todas las modalidades de falsificación. El auditor ya no puede limitarse a preguntar si un documento es coherente con la transacción; debe establecer quién lo creó, cómo fue transmitido y qué controles protegieron su integridad. Esta ampliación no reemplaza los criterios tradicionales de suficiencia y pertinencia, sino que añade la autenticidad como condición previa.

La auditoría debe examinar si las autorizaciones dependen de señales fáciles de imitar. Schmitt y Flechais (2024) describen cómo la inteligencia artificial explota relaciones de confianza y características personales, mientras que Vecchietti et al. (2025) vinculan la gestión de estos riesgos con la gobernanza organizacional. Una evaluación útil no se limita a comprobar que existe una política, sino que identifica si una voz, un rostro o un correo pueden desplazar los controles formales. El hallazgo más relevante puede no ser la ausencia de tecnología, sino la existencia de excepciones que permiten saltarse los procedimientos cuando la instrucción parece provenir de una autoridad.

Las pruebas de auditoría deberían incluir simulaciones controladas de suplantación. Gong y Li (2024) señalan que las estrategias de defensa deben adaptarse a formas cambiantes de manipulación, mientras que Schmitt y Flechais (2024) destacan la importancia de respuestas que integren personas, procesos y tecnología. Un ejercicio autorizado puede revelar si el personal verifica los cambios bancarios, respeta los límites de aprobación y reporta solicitudes sospechosas. Su propósito no debe ser ridiculizar a quien fue engañado, sino detectar debilidades que un delincuente real podría explotar.

La documentación del juicio adquiere mayor importancia cuando existen dudas sobre la autenticidad. Babaei et al. (2025) advierten que las soluciones de reconocimiento no ofrecen resultados infalibles, mientras que Sandoval et al. (2024) muestran que la presencia de contenidos manipulados dificulta la valoración probatoria. El auditor debe dejar constancia de los procedimientos realizados, las fuentes consultadas, las limitaciones técnicas y las razones que sostienen su conclusión. Una decisión no documentada es difícil de defender cuando posteriormente aparece una versión contradictoria de la misma evidencia.

3. Conclusiones y reflexiones

Los *deepfakes* financieros representan una evolución cualitativa del fraude corporativo porque ya no imitan únicamente documentos o mensajes: reproducen a las personas cuya autoridad activa las decisiones. Una voz conocida, una imagen coherente o una conversación aparentemente espontánea pueden ser construidas para provocar una acción real. La consecuencia principal es que la presencia audiovisual ha dejado de ser equivalente a la identidad.

El control interno debe adaptarse a esta pérdida de certeza. La respuesta no consiste en eliminar la confianza de las relaciones empresariales, sino en separar la confianza interpersonal de la aprobación financiera. Una instrucción puede ser creíble y, aun así, requerir credenciales, registro, corroboración y revisión independiente antes de generar efectos sobre los recursos de la entidad.

La segregación de funciones adquiere un significado adicional frente a las falsificaciones sintéticas. Tradicionalmente se utilizaba para reducir la posibilidad de que una persona cometiera y ocultara un fraude. Ahora también debe impedir que alguien engañado pueda completar toda la operación. Este enfoque reconoce la vulnerabilidad humana sin responsabilizar injustamente al trabajador que fue objeto de una manipulación sofisticada.

La evidencia electrónica exige una revisión que vaya más allá de su contenido visible. Una grabación puede resultar coherente; una factura puede parecer impecable; una videollamada puede reproducir gestos familiares. Ninguno de estos elementos demuestra por sí mismo el origen del archivo. La procedencia, los metadatos, la custodia y la confirmación externa se convierten en componentes esenciales de su valor probatorio.

Tampoco sería prudente confiar exclusivamente en programas de detección. La carrera entre generación y reconocimiento impide considerar infalible cualquier solución disponible. El detector puede aportar una señal útil, pero la conclusión debe surgir de la combinación entre análisis técnico, conocimiento del proceso, contrastación independiente y juicio profesional documentado.

La respuesta institucional más sólida es aquella que limita las consecuencias incluso cuando la falsificación no fue descubierta a tiempo. Una organización preparada no depende de que todos sus trabajadores identifiquen cada engaño; distribuye responsabilidades, protege los datos maestros, refuerza las operaciones atípicas y mantiene rutas claras para detener desembolsos.

Finalmente, la amenaza más seria no se encuentra solo en la capacidad de fabricar una voz o un rostro, sino en la permanencia de procedimientos diseñados para una época en la que esos elementos resultaban difíciles de reproducir. El contador y el auditor están llamados a reconstruir la confianza sobre bases verificables. En un entorno donde la apariencia puede crearse, la trazabilidad, la independencia del control y la responsabilidad definida se convierten en las formas más sólidas de autenticidad.

Referencias

Altuncu, E., Franqueira, V. N. L., Li, S., & Alsaadi, H. (2024). Deepfake: Definitions, performance metrics and standards, datasets, and a meta-review. *Frontiers in Big Data, 7*, 1402057. https://doi.org/10.3389/fdata.2024.1402057

Babaei, R., Chen, B., Dehghantanha, A., & Zarrinkalam, F. (2025). Generative artificial intelligence and the evolving challenge of deepfake detection: A systematic analysis. *Journal of Sensor and Actuator Networks, 14*(1), 17. https://doi.org/10.3390/jsan14010017

Gong, L. Y., & Li, X. J. (2024). A contemporary survey on deepfake detection: Datasets, algorithms, and challenges. *Electronics, 13*(3), 585. https://doi.org/10.3390/electronics13030585

Sandoval, M. P., De Almeida Vau, M., Solaas, J., & Rodrigues, L. (2024). Threat of deepfakes to the criminal justice system: A systematic review. *Crime Science, 13*, 41. https://doi.org/10.1186/s40163-024-00239-1

Schmitt, M., & Flechais, I. (2024). Digital deception: Generative artificial intelligence in social engineering and phishing. *Artificial Intelligence Review, 57*, 324. https://doi.org/10.1007/s10462-024-10973-2

Vecchietti, G., Liyanaarachchi, G., & Viglia, G. (2025). Managing deepfakes with artificial intelligence: Introducing the business privacy calculus. *Journal of Business Research, 186*, 115010. https://doi.org/10.1016/j.jbusres.2024.115010