SOL21.org
STANDARD OPEN LITERATURE
SCIENTIFIC JOURNAL
ISSN 3086-089X

v.2, n.1, 2026

REVISTA CIENTÍFICA SOL21
STANDARD OPEN LITERATURE - ISSN 3086-089X

INGENIERÍA SOCIAL POTENCIADA POR INTELIGENCIA ARTIFICIAL GENERATIVA: ESTUDIO DE CASO DEL FRAUDE CORPORATIVO MEDIADO POR DEEPFAKE

ENGENHARIA SOCIAL POTENCIALIZADA POR INTELIGÊNCIA ARTIFICIAL GENERATIVA: ESTUDO DE CASO DA FRAUDE CORPORATIVA MEDIADA POR DEEPFAKE

AI-ENHANCED SOCIAL ENGINEERING: A CASE STUDY OF DEEPFAKE-MEDIATED CORPORATE FRAUD

Marcelo de Melo Marcon[1]

e2137
https://doi.org/10.67553/sol21.v.2i1.37

PUBLICADO: 10/2026

RESUMEN

La convergencia entre inteligencia artificial generativa e ingeniería social ha redefinido el perfil de riesgo de las organizaciones, desplazando el vector de ataque desde el perímetro técnico hacia los procesos de autorización y los mecanismos humanos de verificación de identidad. Este artículo tiene por objetivo inferir, a partir de evidencias públicas, las brechas de control organizacional explotadas por esos ataques y proponer controles compensatorios aplicables. Se adopta la estrategia de estudio de caso único holístico, de naturaleza cualitativa y base documental, cuya unidad de análisis es el incidente ocurrido en enero de 2024 en la oficina de Hong Kong de una empresa multinacional británica de ingeniería, en el que un profesional del área financiera ejecutó quince transferencias por un total aproximado de 25,6 millones de dólares tras participar en una videoconferencia en la que todos los demás participantes eran síntesis audiovisuales generadas por inteligencia artificial. Las evidencias se obtuvieron exclusivamente de fuentes públicas y se sometieron a triangulación, con indicación de las fuentes que sustentan cada hecho. El análisis mapea la cadena de ataque contra el MITRE ATT&CK, versión 19.2, y confronta el incidente con la ISO/IEC 27001:2022 y el NIST Cybersecurity Framework 2.0. No se reportó compromiso tecnológico de los sistemas, y las evidencias apuntan a fragilidades de naturaleza procedimental, entre ellas la aceptación de la evidencia sensorial como prueba de identidad. Se propone reposicionar la videoconferencia como canal no autenticado.

PALABRAS CLAVE: Ingeniería social. Deepfake. Inteligencia artificial generativa. Fraude corporativo. Gobernanza de seguridad de la información.

RESUMO

The convergence of generative artificial intelligence and social engineering has redefined organizational risk profiles, shifting the attack vector from the technical perimeter toward authorization processes and human identity-verification mechanisms. This article infers, from public evidence, the organizational control gaps exploited by AI-assisted social engineering attacks and proposes applicable compensating controls. A holistic single-case study is adopted, taking as its unit of analysis the January 2024 incident at the Hong Kong office of a British multinational engineering firm, in which a finance professional executed fifteen transfers totaling approximately USD 25.6 million after attending a video conference in which every other participant was an AI-generated audiovisual synthesis. Evidence was drawn exclusively from public sources and triangulated, with the sources supporting each fact identified. The analysis maps the attack chain against MITRE

[1] Doutor Honoris Causa em Ciências Exatas pela World Christian University (WCU), Mestre em Direção Estratégica em Tecnologias da Informação pela Universidad Europea del Atlántico, pós-graduado em MBIS – Segurança da Informação e em Engenharia de Redes e Sistemas de Telecomunicações. Possui mais de 20 anos de experiência profissional em Tecnologia da Informação e Comunicação (TIC), com atuação em ambientes críticos nos setores de saúde, governo e telecomunicações. Sua experiência abrange Segurança da Informação, Cibersegurança, Proteção de Dados, Gestão de Riscos, Datacenters, Virtualização, Linux, Redes, Governança, Continuidade de Negócios e Melhoria de Processos. É instrutor da Escola Superior de Guerra – ADESG/Campinas, membro do Comitê Avançado e Altos Estudos de Tecnologia e Inovação da Digital Law Academy, Diretor do Comitê de Segurança da APDADOS e Imortal da Academia Brasileira de Ciências, Artes, História e Literatura (ABRASCI).

Este artículo se publica en acceso abierto (Open Access) bajo licencia Creative Commons Attribution 4.0 Internacional (CC-BY), que permite uso, distribución y reproducción irrestrictos en cualquier medio, siempre que el autor original y la fuente sean acreditados.

REVISTA CIENTÍFICA SOL21
STANDARD OPEN LITERATURE - ISSN 3086-089X

ATT&CK techniques, version 19.2, and confronts the incident with ISO/IEC 27001:2022 and the NIST Cybersecurity Framework 2.0. No technological compromise of the organization's systems was reported, and the evidence examined points to procedural weaknesses, among them the acceptance of sensory evidence as proof of identity. The article proposes repositioning video conferencing as an unauthenticated channel.

PALAVRAS-CHAVE: Engenharia social. Deepfake. Inteligência artificial generativa. Fraude corporativa. Governança de segurança da informação.

ABSTRACT
La convergencia entre inteligencia artificial generativa e ingeniería social ha redefinido el perfil de riesgo de las organizaciones, desplazando el vector de ataque desde el perímetro técnico hacia los procesos de autorización y los mecanismos humanos de verificación de identidad. Este artículo tiene por objetivo inferir, a partir de evidencias públicas, las brechas de control organizacional explotadas por ataques de ingeniería social asistidos por inteligencia artificial generativa y proponer controles compensatorios aplicables. Se adopta la estrategia de estudio de caso único holístico, cuya unidad de análisis es el incidente ocurrido en enero de 2024 en la oficina de Hong Kong de una empresa multinacional británica de ingeniería, en el que un profesional del área financiera ejecutó quince transferencias por un total aproximado de 25,6 millones de dólares tras participar en una videoconferencia en la que todos los demás participantes eran síntesis audiovisuales generadas por inteligencia artificial. Las evidencias se obtuvieron exclusivamente de fuentes públicas y se sometieron a triangulación. El análisis mapea la cadena de ataque contra el MITRE ATT&CK, versión 19.2, y confronta el incidente con la ISO/IEC 27001:2022 y el NIST Cybersecurity Framework 2.0. No se reportó compromiso tecnológico de los sistemas, y las evidencias apuntan a fragilidades de naturaleza procedimental. Se propone reposicionar la videoconferencia como canal no autenticado.

KEYWORDS: Social engineering. Deepfake. Generative artificial intelligence. Corporate fraud. Information security governance.

INTRODUCCIÓN

Durante décadas, la doctrina de seguridad de la información se organizó en torno a la defensa del perímetro tecnológico, arquitectura que presupone que el ataque deja rastro técnico. Cuando el adversario no invade sistema alguno y obtiene el mismo resultado mediante la manipulación de un proceso autorizativo legítimo, los controles perimetrales se vuelven irrelevantes por construcción. La ingeniería social siempre representó esa excepción. Lo que cambió a partir de 2023 no fue la lógica del ataque, sino su escala de credibilidad: los modelos generativos de síntesis audiovisual se volvieron accesibles, baratos y capaces de operar en tiempo real. La evidencia sensorial —ver el rostro de un colega y oír su voz— dejó de constituir prueba de identidad, aunque siga operando como tal en la mayoría de los procesos organizacionales.

Los datos agregados confirman la materialidad del fenómeno. El Internet Crime Complaint Center del Federal Bureau of Investigation registró, en 2025, más de un millón de reclamaciones y pérdidas superiores a 20.800 millones de dólares, de las cuales cerca de 3.000 millones se atribuyeron al compromiso del correo corporativo; en el mismo informe, la inteligencia artificial se incorporó por primera vez como descriptor formal de modalidad criminal, con 22.364 reclamaciones

Este artículo se publica en acceso abierto (Open Access) bajo licencia Creative Commons Attribution 4.0 Internacional (CC-BY), que permite uso, distribución y reproducción irrestrictos en cualquier medio, siempre que el autor original y la fuente sean acreditados.

y cerca de 893 millones de dólares (Federal Bureau of Investigation [FBI], 2026). Proyecciones del sector financiero estiman que los fraudes habilitados por IA generativa alcancen 40.000 millones de dólares anuales en los Estados Unidos hacia 2027, frente a 12.300 millones en 2023 (Deloitte, 2024).

En ese contexto, el incidente sufrido por una empresa multinacional británica de ingeniería en enero de 2024 constituye un caso paradigmático. Un profesional del área financiera de la oficina de Hong Kong ejecutó quince transferencias por un total aproximado de 25,6 millones de dólares estadounidenses tras participar en una videoconferencia en la que creía dialogar con el director financiero global y otros colegas, todas representaciones sintéticas generadas a partir de material audiovisual públicamente disponible. Ningún sistema fue comprometido y ningún dato fue exfiltrado (Magramo, 2024; Prakash, 2024). El caso es relevante porque la ausencia de compromiso técnico aísla la variable de interés: no hay malware que analizar ni vulnerabilidad que corregir. Resta la arquitectura de autorización financiera y los supuestos de verificación de identidad en ella incorporados.

Se delimita la siguiente pregunta de investigación: ¿qué brechas de control organizacional son explotadas por ataques de ingeniería social asistidos por inteligencia artificial generativa y qué controles son capaces de mitigarlas? El objetivo general es identificar dichas brechas a partir del análisis sistemático de un caso real, confrontándolo con marcos consolidados. Como objetivos específicos: (i) reconstruir la cadena de ataque a partir de evidencias públicas trianguladas; (ii) mapear las técnicas según el MITRE ATT&CK; (iii) identificar las brechas a la luz del Anexo A de la ISO/IEC 27001:2022; y (iv) proponer controles compensatorios organizados por las funciones del NIST Cybersecurity Framework 2.0.

La justificación se apoya en tres razones. En el conjunto de literatura examinado por los autores, la producción sobre deepfakes se concentra en la base técnica de generación y detección (Goodfellow et al., 2014; Mirsky y Lee, 2021) y en implicaciones jurídicas y político-electorales (Westerlund, 2019; Chesney y Citron, 2019), siendo menos frecuente el examen del fenómeno como problema de control organizacional, delimitación que se refiere al corpus consultado y no a una revisión sistemática. En segundo lugar, la reconstrucción documentada de un incidente de alto impacto ofrece insumo directo para la revisión de políticas de autorización financiera. En tercero, la erosión de la confianza en la evidencia audiovisual afecta los mecanismos de verificación sobre los que se asientan relaciones contractuales e institucionales.

1 - MARCO TEÓRICO

1.1 INGENIERÍA SOCIAL Y LOS LÍMITES DE LA CONCIENCIACIÓN

La ingeniería social designa las técnicas destinadas a inducir a las personas a ejecutar acciones o revelar informaciones que comprometen la seguridad, mediante la manipulación de la confianza y no de la tecnología. Mitnick y Simon (2002) demostraron que el elemento humano es el

Este artículo se publica en acceso abierto (Open Access) bajo licencia Creative Commons Attribution 4.0 Internacional (CC-BY), que permite uso, distribución y reproducción irrestrictos en cualquier medio, siempre que el autor original y la fuente sean acreditados.

vector de menor resistencia en cualquier arquitectura de seguridad, por operar según heurísticas sociales que no pueden desactivarse sin inviabilizar la cooperación organizacional. Hadnagy (2018) sistematiza el ataque en fases —recolección, pretexto, vínculo, explotación y ejecución—, resaltando que la calidad del pretexto depende del reconocimiento previo.

El fundamento psicológico de esa eficacia fue mapeado por Cialdini (2009) en seis principios de influencia, de los cuales tres operan con particular intensidad en contextos corporativos: autoridad, escasez y prueba social. La concienciación aislada, sin embargo, presenta eficacia limitada: las capacitaciones basadas en el reconocimiento de indicios —errores gramaticales, dominios sospechosos, remitentes anómalos— presuponen que el ataque contenga imperfecciones detectables. Cuando la síntesis generativa elimina esos indicios, el modelo de defensa fundado en la percepción individual pierde sustento, restando solo los controles de proceso, lo que desplaza el problema del dominio de la concienciación al de la gobernanza.

1.2 DEEPFAKES E INTELIGENCIA ARTIFICIAL GENERATIVA

El término deepfake designa contenido audiovisual sintético producido por arquitecturas de aprendizaje profundo, típicamente con el propósito de sustituir o recrear la identidad de un individuo. La base técnica se remonta a las redes generativas adversarias propuestas por Goodfellow et al. (2014), en las que dos redes neuronales compiten —una generando muestras, otra discriminando entre sintéticas y reales—, elevando progresivamente la fidelidad de la síntesis. Mirsky y Lee (2021) advierten que los métodos de detección se degradan significativamente cuando se aplican a datos fuera de la distribución de entrenamiento; la carrera entre generación y detección favorece estructuralmente al generador, pues cualquier detector publicado se convierte en blanco de optimización adversaria. Westerlund (2019) argumenta que el impacto del deepfake deriva menos de su calidad absoluta que del colapso de la barrera de entrada, y Chesney y Citron (2019) formulan el mismo fenómeno en términos jurídicos, acuñando la noción de dividendo del mentiroso: la difusión de contenido sintético degrada la autoridad probatoria de todo registro audiovisual, incluidos los auténticos. Importa retener tres atributos: la síntesis en tiempo real se volvió viable en plataformas convencionales; el material de entrenamiento se obtiene de fuentes abiertas; y la detección automatizada no constituye hoy un control confiable.

1.3 FRAUDE DEL CEO Y REFERENCIALES DE CONTROL

El compromiso del correo corporativo consiste en la inducción de una transferencia financiera mediante la personificación de una figura de autoridad interna o de un socio comercial. Según el FBI (2026), la categoría representó 3.046 millones de dólares en pérdidas reportadas en 2025, en 24.768 reclamaciones, siendo la segunda mayor por valor; el informe anterior registró 2.770 millones en 21.442 reclamaciones (FBI, 2025). La estructura canónica comprende el reconocimiento

Este artículo se publica en acceso abierto (Open Access) bajo licencia Creative Commons Attribution 4.0 Internacional (CC-BY), que permite uso, distribución y reproducción irrestrictos en cualquier medio, siempre que el autor original y la fuente sean acreditados.

de la jerarquía y de los flujos de pago, la personificación por canal asíncrono, el pretexto de confidencialidad y urgencia, y el direccionamiento a cuentas del adversario. La defensa consagrada se apoya en la verificación fuera de banda y en la segregación de funciones. La innovación del caso está en la extensión del ataque a un canal síncrono y audiovisual: cuando el canal considerado más confiable pasa a ser falsificable, el propio control compensatorio queda neutralizado.

El análisis recurre a tres instrumentos consolidados. El MITRE ATT&CK es una base de conocimiento de tácticas y técnicas adversarias observadas empíricamente; a partir de 2023 pasó a contemplar técnicas de personificación y de robo financiero, volviéndose aplicable a incidentes sin compromiso de sistemas. En la versión 19.2, consultada en octubre de 2026 y adoptada en este trabajo, esas técnicas corresponden a T1684.001 (Impersonation), subtécnica de T1684 (Social Engineering), y a T1657 (Financial Theft) (MITRE Corporation, 2026). La ISO/IEC 27001:2022 establece requisitos para sistemas de gestión de seguridad de la información (International Organization for Standardization [ISO], 2022). Y el NIST Cybersecurity Framework 2.0 organiza resultados en seis funciones, siendo relevante la introducción de la función Gobernar, que explicita que las decisiones de riesgo y las políticas anteceden a los controles técnicos (National Institute of Standards and Technology [NIST], 2024).

2 - METODOLOGÍA

2.1 DISEÑO DE LA INVESTIGACIÓN Y UNIDAD DE ANÁLISIS

La investigación es cualitativa en cuanto al enfoque, descriptiva y explicativa en cuanto a los objetivos, y documental en cuanto a los procedimientos. Se adopta la estrategia de estudio de caso único holístico, conforme a Yin (2018), justificada por dos criterios: el carácter revelador del fenómeno, por tratarse de uno de los primeros incidentes de fraude mediado por síntesis audiovisual en tiempo real documentado públicamente y con atribución confirmada por la víctima; y la magnitud del impacto, que asegura una densidad documental poco común. El mapeo de técnicas adopta la versión 19.2 del MITRE ATT&CK for Enterprise, consultada en octubre de 2026.

La unidad de análisis es el incidente ocurrido en enero de 2024 en la oficina de Hong Kong de la organización, delimitado entre la preparación del ataque y la divulgación pública de la identidad de la víctima, en mayo de 2024. Quedan fuera del alcance los desdoblamientos investigativos posteriores.

2.2 FUENTES DE EVIDENCIA Y TRIANGULACIÓN

Las evidencias se obtuvieron exclusivamente de fuentes públicas, lo que dispensa la sumisión a un Comité de Ética en Investigación, dado que no hubo recolección con seres humanos ni acceso a informaciones no divulgadas. Se consultaron tres categorías: reportajes de medios periodísticos de referencia, en los cuales se encuentran reproducidos el comunicado de la

Este artículo se publica en acceso abierto (Open Access) bajo licencia Creative Commons Attribution 4.0 Internacional (CC-BY), que permite uso, distribución y reproducción irrestrictos en cualquier medio, siempre que el autor original y la fuente sean acreditados.

organización víctima y las informaciones divulgadas por la policía de Hong Kong; informes de autoridades y organismos sectoriales, empleados para la caracterización del contexto y no del incidente; y literatura revisada por pares. Regístrese que el comunicado corporativo y las declaraciones policiales no se obtuvieron en documentos autónomos, sino en los reportajes que los reproducen, razón por la cual se citan por intermedio de ellos.

Se aplicó el criterio de triangulación de Yin (2018): solo se incorporaron los elementos fácticos confirmados por al menos dos fuentes independientes. Se consideraron independientes las fuentes de propiedad editorial distinta, con autoría propia y cobertura no derivada de un mismo despacho de agencia, condición satisfecha por Magramo (2024), de CNN, y Prakash (2024), de Fortune. La columna de fuentes de la Tabla 1 indica, por evento, qué documentos lo sustentan. Informaciones de proveedores de soluciones de seguridad no se usaron como base fáctica. Por decisión de los autores, la organización víctima no se identifica nominalmente en el cuerpo de este artículo, aunque los documentos periodísticos la identifiquen y sus títulos se reproduzcan en las referencias conforme fueron publicados.

2.3 LIMITACIONES DEL ESTUDIO

No hubo acceso a documentación interna, registros de auditoría o informes forenses. En consecuencia, las brechas presentadas en la sección 3.3 constituyen inferencias fundamentadas en la descripción pública del incidente, y no constataciones de auditoría: las fuentes disponibles no permiten distinguir entre la inexistencia de un control, su existencia con elusión, la concesión de una excepción operativa u otra circunstancia no documentada. Esa salvedad vale para todo el texto. Se registran otras tres limitaciones: el estudio de caso único permite generalización analítica, pero no estadística (Yin, 2018); hay posibilidad de sesgo de supervivencia; y la capacidad de síntesis audiovisual evolucionó desde la ocurrencia, de modo que las recomendaciones deben leerse como orientaciones de proceso.

2.4 DECLARACIÓN DEL USO DE INTELIGENCIA ARTIFICIAL

Hubo uso de IA: Los autores declaran que utilizaron el asistente Claude, de Anthropic, modelo Claude Opus 5, accedido entre septiembre y octubre de 2026, en el relevamiento y cribado de fuentes, la definición del diseño, la redacción, la reconstrucción de la cadena de ataque, la elaboración del mapeo contra el MITRE ATT&CK, la identificación de las brechas, la formulación de los controles compensatorios y el formateo. Regístrese, para precisión de la declaración, que esa participación no se limitó a la asistencia textual: abarcó la organización analítica del material y la proposición inicial del diseño, del mapeo, de las brechas y de los controles, correspondiendo a los autores la decisión sobre cada uno de esos elementos, su verificación y su validación. El uso afectó la totalidad del texto, incluidas las Tablas 1 a 4 y las referencias, y abarca la revisión realizada en respuesta al dictamen. Son de responsabilidad exclusiva de los autores la definición del tema y del

Este artículo se publica en acceso abierto (Open Access) bajo licencia Creative Commons Attribution 4.0 Internacional (CC-BY), que permite uso, distribución y reproducción irrestrictos en cualquier medio, siempre que el autor original y la fuente sean acreditados.

recorte, la elección del caso, las decisiones sobre el diseño, los criterios de triangulación, la decisión de no identificar nominalmente a la organización y la validación final. Toda la información sobre el incidente proviene de reportajes periodísticos, en los cuales se reproducen el comunicado de la organización y las declaraciones policiales; no se suministraron al sistema datos internos, confidenciales u obtenidos por vía no pública, y el profesional involucrado no es identificado ni caracterizado, razón por la cual no hubo tratamiento de datos sujetos a la Ley n.º 13.709/2018 de Brasil. Los prompts y registros metodológicamente relevantes están archivados por los autores y disponibles a la editoría mediante solicitud. Las salidas fueron revisadas mediante el cotejo de cada referencia contra la fuente primaria, de los identificadores digitales de objeto, de los códigos del MITRE ATT&CK contra la versión 19.2 y de los identificadores de control de la ISO/IEC 27001:2022. Ese cotejo, rehecho en esta revisión, llevó a la corrección de dos encuadres del mapeo, conforme a la nota de la Tabla 2. Ningún dato, cita o conclusión fue aceptado sin verificación humana.

3 - RESULTADOS Y DISCUSIÓN

3.1 EL CASO Y LA RECONSTRUCCIÓN DE LA CADENA DE ATAQUE

La organización es una empresa multinacional británica de consultoría en ingeniería, con sede en Londres, que en la época declaraba operar con cerca de dieciocho mil quinientos profesionales en treinta y cuatro oficinas (Magramo, 2024). El perfil es relevante por dos razones: la dispersión geográfica reduce la familiaridad interpersonal entre matriz y unidades, volviendo la comunicación mediada por tecnología la norma; y la alta exposición pública de los ejecutivos pone a disposición material audiovisual en fuentes abiertas. La Tabla 1 consolida la secuencia de eventos.

Tabla 1

Cronología del incidente

Momento	Evento	Fuentes
Período previo	Recolección de material audiovisual público de ejecutivos, obtenido de grabaciones de eventos, entrevistas y reuniones corporativas divulgadas, y utilizado para el entrenamiento de los modelos de síntesis.	Magramo (2024); Prakash (2024)
Enero de 2024	Recepción, por un profesional del área financiera de la oficina de Hong Kong, de un mensaje electrónico atribuido al director financiero con sede en el Reino Unido, solicitando una transacción confidencial. El profesional manifiesta sospecha e identifica indicios compatibles con phishing.	Magramo (2024); Prakash (2024)
Enero de 2024	Realización de una videoconferencia con participación aparente del director financiero y de otros colegas reconocibles, todos sintéticamente generados. La solicitud es reiterada y confirmada en el canal síncrono.	Magramo (2024); Prakash (2024)
Enero de 2024	Ejecución de quince transferencias, por un total de cerca de doscientos millones de dólares de Hong Kong, a cinco cuentas distintas.	Magramo (2024); Prakash (2024)
Febrero de 2024	Descubrimiento del fraude en un contacto posterior del profesional con la sede, comunicación a la policía de Hong Kong y divulgación del caso por la policía, sin identificación nominal de la víctima.	Magramo (2024); Prakash (2024)

Este artículo se publica en acceso abierto (Open Access) bajo licencia Creative Commons Attribution 4.0 Internacional (CC-BY), que permite uso, distribución y reproducción irrestrictos en cualquier medio, siempre que el autor original y la fuente sean acreditados.

REVISTA CIENTÍFICA SOL21
STANDARD OPEN LITERATURE - ISSN 3086-089X

Momento	Evento	Fuentes
Mayo de 2024	Confirmación pública por la organización, con la aclaración de que ningún sistema interno fue comprometido y ningún dato fue afectado.	Comunicado de la organización reproducido en Magramo (2024) y en Prakash (2024)

Nota. Elaborado por los autores. La columna de fuentes indica los documentos que sustentan cada evento; todos los eventos relacionados cuentan con dos fuentes de propiedad editorial distinta, conforme al criterio de independencia definido en la subsección 2.2. Los datos de contexto de Deloitte (2024) y del FBI (2026) no sustentan hechos del incidente y por ello no figuran en esta tabla.

La reconstrucción permite identificar cinco fases. El reconocimiento consistió en la identificación de la estructura jerárquica, del flujo de autorización y del profesional competente, y en la recolección del material audiovisual: fase pasiva, sobre fuentes abiertas, indetectable por la organización objetivo. El desarrollo de recursos involucró cinco cuentas receptoras, lo que fracciona los valores, reduce la probabilidad de activación de umbrales antifraude y distribuye el riesgo de bloqueo. El contacto inicial ocurrió por canal asíncrono y falló: el profesional identificó la anomalía y no ejecutó la solicitud, de modo que el control humano funcionó conforme a lo esperado.

La superación de la resistencia es el elemento distintivo del caso. Ante la sospecha manifestada, el adversario no abandonó el objetivo ni intensificó la presión por el mismo canal, sino que migró a un canal de mayor credibilidad percibida. La videoconferencia con múltiples participantes reconocibles movilizó tres principios de influencia: autoridad, prueba social y escasez. Es precisamente el mecanismo de verificación adoptado por el profesional el que fue convertido en instrumento de persuasión. El descubrimiento ocurrió por iniciativa del propio profesional, y no por un control automatizado.

3.2 MAPEO SEGÚN EL MITRE ATT&CK

Tabla 2

Mapeo de la cadena de ataque según el MITRE ATT&CK

Fase	Técnica	Evidencia en el caso
Reconocimiento	T1589 — Gather Victim Identity Information	Identificación de ejecutivos y del profesional responsable de la ejecución de pagos.
Reconocimiento	T1593 — Search Open Websites/Domains	Obtención, en sitios corporativos y redes sociales, del material audiovisual público utilizado en el entrenamiento de los modelos de síntesis.
Acceso inicial	T1566 — Phishing	Mensaje electrónico atribuido al director financiero solicitando una transacción confidencial.
Stealth	T1684.001 — Impersonation	Personificación audiovisual en tiempo real del director financiero y de colegas en videoconferencia.
Impacto	T1657 — Financial Theft	Ejecución de quince transferencias a cinco cuentas bancarias previamente constituidas y controladas por el adversario.

Nota. Elaborado por los autores con base en MITRE Corporation (2026), versión 19.2 de la taxonomía, consultada en octubre de 2026, y en las evidencias del caso. Dos encuadres de la versión anterior de este mapeo fueron corregidos en la revisión. La personificación constaba como T1656; en la versión 19.2 esa técnica integra T1684 (Social Engineering) como subtécnica T1684.001, bajo la táctica Stealth, antes denominada Evasión de Defensas. Y la constitución de las cuentas bancarias receptoras constaba como T1585 (Establish Accounts), encuadre retirado por no corresponder a la definición de la técnica, restringida a cuentas en servicios en línea destinadas a la construcción de persona; la evidencia fue reasignada a T1657, cuya descripción contempla el envío de recursos a cuentas controladas por el adversario.

Este artículo se publica en acceso abierto (Open Access) bajo licencia Creative Commons Attribution 4.0 Internacional (CC-BY), que permite uso, distribución y reproducción irrestrictos en cualquier medio, siempre que el autor original y la fuente sean acreditados.

Tres observaciones se derivan del mapeo. La primera es la ausencia de técnicas de las tácticas de ejecución, persistencia, movimiento lateral y exfiltración: la cadena transita del acceso inicial a la personificación y al impacto, sin interacción con los sistemas de la organización. La segunda es que la técnica decisiva, T1684.001, no posee contramedida técnica convencional y la mitigación prevista es procedimental. La tercera es que la constitución previa de cuentas bancarias receptoras, aunque documentada en el caso, no encuentra técnica correspondiente en la matriz Enterprise, construida para describir la intrusión en sistemas y no la cadena financiera que monetiza el fraude: límite de la taxonomía que el propio encuadre de este caso vuelve visible.

3.3 BRECHAS DE CONTROL A LA LUZ DE LA ISO/IEC 27001:2022

Tabla 3

Brechas de control inferidas según el Anexo A de la ISO/IEC 27001:2022

Control	Resultado esperado	Brecha inferida
5.3 — Segregación de funciones	Imposibilidad de que una única persona inicie y concluya una transacción de alto valor.	La ejecución de quince transferencias de valor agregado elevado por un único profesional es compatible con la hipótesis de ausencia de segregación vinculante o de su suspensión mediante invocación de confidencialidad. Las fuentes públicas no permiten determinar el flujo interno de aprobación ni distinguir entre inexistencia del control, su elusión y una eventual excepción autorizada.
5.14 — Transferencia de información	Definición de canales autorizados y de reglas de verificación para comunicaciones sensibles.	La aceptación de la videoconferencia como confirmación suficiente sugiere la ausencia de exigencia de un canal independiente previamente establecido.
5.15 y 5.17 — Control de acceso y autenticación	Verificación de identidad basada en factores no replicables.	La verificación se apoyó en el reconocimiento sensorial de rostro y voz, atributos públicamente disponibles y sintéticamente reproducibles.
8.16 — Actividades de monitoreo	Detección de patrones transaccionales anómalos en tiempo hábil.	La detección ocurrió por iniciativa humana posterior, sin registro público de alerta por fraccionamiento o destino atípico.

Nota. Elaborado por los autores con base en International Organization for Standardization (2022) y en las evidencias públicas del caso. Las brechas son inferencias de los autores a partir del desenlace documentado, y no constataciones de auditoría.

Se suman a las brechas de la tabla otras dos, igualmente inferidas: el control 5.7, de inteligencia de amenazas, aparentemente no contemplaba la síntesis audiovisual en tiempo real como vector previsto en el proceso de pago; y el control 6.3, de concienciación, se orientaba a la identificación de indicios en canales asíncronos, sin cobertura de personificación síncrona. La brecha más consecuente es la relativa a la verificación de identidad: los demás controles presuponen un mecanismo confiable para establecer quién está del otro lado del canal, y cuando ese mecanismo es la percepción sensorial, vuelta reproducible, la cadena de controles dependientes queda comprometida. La expresión brecha inferida se usa en sentido estricto: no se afirma que la organización no poseyera políticas formales, sino que el desenlace documentado es más compatible con la hipótesis de una falla de eficacia de esos controles.

Este artículo se publica en acceso abierto (Open Access) bajo licencia Creative Commons Attribution 4.0 Internacional (CC-BY), que permite uso, distribución y reproducción irrestrictos en cualquier medio, siempre que el autor original y la fuente sean acreditados.

3.4 LA OBSOLESCENCIA DE LA EVIDENCIA SENSORIAL COMO PRUEBA DE IDENTIDAD

La contribución analítica central se refiere al supuesto implícito que sustenta los procesos autorizativos organizacionales. La proposición hablé con la persona operó históricamente como patrón de oro de confirmación, sin haber sido jamás explicitada. El caso ofrece evidencia de que ese supuesto no se sostiene ante la síntesis audiovisual en tiempo real, y de forma aguda, porque el profesional no actuó con negligencia: identificó el indicio en el canal asíncrono, manifestó sospecha y recurrió a un mecanismo de verificación. Lo que falló no fue el juicio individual, sino la validez del instrumento disponible —distinción relevante, pues atribuir el incidente a error humano conduce a respuestas centradas en la capacitación, cuando la corrección necesaria es estructural—

Se formula, a partir de ese análisis, la proposición que sintetiza el resultado del estudio: en entornos en los que la síntesis audiovisual en tiempo real es técnicamente accesible, la videoconferencia debería clasificarse como canal no autenticado, equiparándose, para fines de control, al mensaje electrónico y a la llamada telefónica. Se trata de una proposición analítica y de una recomendación de gobernanza derivadas de este caso y de la literatura, y no de una regla cuya validez universal un estudio de caso único pueda demostrar. De ella se deriva que la confirmación por videoconferencia no debería aceptarse como satisfacción del requisito de verificación fuera de banda. La proposición contraría la orientación, común en materiales de concienciación, de identificar deepfakes solicitando gestos al interlocutor: recomendación que transfiere al individuo la carga de una pericia que no le compete y cuya validez es decreciente (Mirsky y Lee, 2021).

3.5 CONTROLES COMPENSATORIOS PROPUESTOS

Tabla 4

Controles compensatorios propuestos por función del NIST CSF 2.0

Función	Control propuesto	Efecto esperado
Gobernar	Clasificar formalmente la videoconferencia como canal no autenticado en la política de autorización financiera.	Explicita el supuesto de confiabilidad y vuelve auditable la decisión de riesgo.
Gobernar	Establecer que las solicitudes de sigilo interno no suspenden los controles de autorización.	Reduce la eficacia del pretexto de confidencialidad, elemento recurrente de la modalidad.
Proteger	Exigir verificación fuera de banda por canal previamente registrado, iniciada por el verificador, para transacciones por encima de un umbral definido.	Dificulta que el adversario controle el canal de confirmación.
Proteger	Instituir segregación de funciones vinculante, con aprobación por dos responsables independientes.	Eleva el costo del ataque, al exigir el compromiso simultáneo de dos decisores.
Proteger	Adoptar una palabra clave o un desafío compartido, definido previamente y por canal distinto, para transacciones excepcionales.	Introduce un factor de verificación no derivable de material público.
Detectar	Implementar reglas de correlación para fraccionamiento de valores, destinos nuevos y desvío de patrón histórico.	Puede ampliar la probabilidad de interrupción durante la ventana de ejecución.

Este artículo se publica en acceso abierto (Open Access) bajo licencia Creative Commons Attribution 4.0 Internacional (CC-BY), que permite uso, distribución y reproducción irrestrictos en cualquier medio, siempre que el autor original y la fuente sean acreditados.

REVISTA CIENTÍFICA SOL21
STANDARD OPEN LITERATURE - ISSN 3086-089X

Función	Control propuesto	Efecto esperado
Responder	Activar el canal de recuperación bancaria en el menor plazo posible, según meta definida en la política de la organización, e instituir una política de no sanción para la comunicación oportuna de sospecha o error.	Puede contribuir a reducir el riesgo de dispersión internacional de los valores y la latencia entre sospecha y reporte.

Nota. Elaborado por los autores con base en National Institute of Standards and Technology (2024), International Organization for Standardization (2022) y en el análisis del caso. Los efectos indicados son proposiciones derivadas del análisis; los controles no fueron implementados ni evaluados empíricamente en este estudio.

Ninguno de los controles propuestos depende de la detección automatizada de contenido sintético: los detectores disponibles se degradan fuera del dominio de entrenamiento y son blanco de optimización adversaria, de modo que fundamentar la defensa en la detección transferiría la seguridad del proceso a una carrera tecnológica desfavorable para el defensor.

3.6 DISCUSIÓN DE LOS RESULTADOS

Los resultados convergen con la literatura en cuanto al papel de los principios de autoridad y escasez, pero agregan un elemento no suficientemente tratado: la movilización de la prueba social por la presencia simultánea de múltiples interlocutores sintéticos, amplificación cualitativa y no cuantitativa. Un interlocutor falsificado exige del objetivo la suspensión de una verificación; varios interlocutores convergentes exigen la suspensión de la hipótesis de que tal coordinación sea factible, hipótesis que el costo marginal casi nulo de la síntesis adicional volvió obsoleta. Los hallazgos dialogan además con Chesney y Citron (2019) sobre la degradación de la autoridad probatoria del registro audiovisual: si la videoconferencia no prueba identidad, prácticas corrientes como aprobaciones remotas y recuperación de credenciales ante el soporte operan sobre un fundamento comprometido.

Desde el punto de vista de la gestión de riesgos, la respuesta adecuada no es necesariamente tecnológica: la asimetría entre generación y detección favorece al adversario, mientras que los controles de proceso permanecen eficaces independientemente de la calidad de la falsificación, porque no dependen de distinguir lo verdadero de lo falso, sino de exigir del adversario el control de canales que él no controla. Se sugiere, por fin, la revisión de los programas de concienciación: los orientados al reconocimiento de indicios tienden a la obsolescencia, mientras que los orientados a la disciplina de procedimiento se muestran más robustos, por exigir que el profesional actúe contra su propia percepción.

4 - CONSIDERACIONES FINALES

Este trabajo examinó, mediante un estudio de caso único holístico de base documental, las brechas de control explotadas por un ataque de ingeniería social asistido por inteligencia artificial generativa, teniendo por objeto el incidente de enero de 2024 en la oficina de Hong Kong de una empresa multinacional británica de ingeniería.

Este artículo se publica en acceso abierto (Open Access) bajo licencia Creative Commons Attribution 4.0 Internacional (CC-BY), que permite uso, distribución y reproducción irrestrictos en cualquier medio, siempre que el autor original y la fuente sean acreditados.

La cadena de ataque fue reconstruida en cinco fases a partir de evidencias públicas trianguladas, evidenciándose que la fase decisiva no fue el contacto inicial, sino la superación de la resistencia legítimamente manifestada por el profesional. El mapeo según el MITRE ATT&CK reveló una cadena atípicamente corta, sin interacción alguna con los sistemas de la organización, lo que indica la inadecuación de los controles perimetrales para esta clase de amenaza. Se infirieron seis brechas de control a partir de las evidencias públicas disponibles, a la luz de la ISO/IEC 27001:2022, de las cuales la más consecuente se refiere a la verificación de identidad, y se propuso un conjunto de controles compensatorios organizados por las funciones del NIST Cybersecurity Framework 2.0, todos independientes de la detección automatizada de contenido sintético.

La contribución teórica consiste en la proposición de que la videoconferencia debería reclasificarse como canal no autenticado, con la consecuencia de que la confirmación por video no satisface el requisito de verificación fuera de banda. La contribución práctica reside en el conjunto de controles propuesto, predominantemente procedimental y de gobernanza, cuya adopción depende sobre todo de una decisión en cuanto a la supresión de excepciones procedimentales; el estudio no evaluó costos, infraestructura ni impacto operativo de su implementación. Se reconocen las limitaciones ya explicitadas, en especial la ausencia de acceso a documentación interna, que confiere a las brechas el estatuto de inferencia.

Como agenda futura, se sugieren cuatro direcciones: un estudio de casos múltiples comparando incidentes de personificación sintética en distintos sectores; una encuesta sobre el grado de adopción de verificación fuera de banda en organizaciones brasileñas; un estudio experimental de la eficacia comparada entre programas de concienciación orientados a indicios y a la disciplina de procedimiento; y un análisis de las implicaciones jurídicas de la personificación sintética en el derecho brasileño, en cuanto a la responsabilidad del empleado que ejecuta una orden falsificada de modo indetectable y al tratamiento de los datos biométricos usados en el entrenamiento de los modelos a la luz de la LGPD.

REFERENCIAS

Chesney, R., & Citron, D. K. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review, 107*(6), 1753–1820. https://doi.org/10.15779/Z38RV0D15J

Cialdini, R. B. (2009). *Influence: Science and practice* (5th ed.). Pearson.

Deloitte. (2024). *Generative AI is expected to magnify the risk of deepfakes and other fraud in banking*. Deloitte Center for Financial Services. https://www.deloitte.com/us/en/insights/industry/financial-services/deepfake-banking-fraud-risk-on-the-rise.html

Federal Bureau of Investigation. (2025). *2024 Internet crime report*. Internet Crime Complaint Center. https://www.ic3.gov/AnnualReport/Reports/2024_IC3Report.pdf

Federal Bureau of Investigation. (2026). *2025 Internet crime report*. Internet Crime Complaint Center. https://www.ic3.gov/AnnualReport/Reports/2025_IC3Report.pdf

Este artículo se publica en acceso abierto (Open Access) bajo licencia Creative Commons Attribution 4.0 Internacional (CC-BY), que permite uso, distribución y reproducción irrestrictos en cualquier medio, siempre que el autor original y la fuente sean acreditados.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (Vol. 27). Curran Associates.

Hadnagy, C. (2018). *Social engineering: The science of human hacking* (2nd ed.). Wiley.

International Organization for Standardization. (2022). *ISO/IEC 27001:2022 — Information security, cybersecurity and privacy protection: Information security management systems: Requirements*.

Lei n. 13.709, de 14 de agosto de 2018. (2018). *Lei Geral de Proteção de Dados Pessoais (LGPD)*. Presidência da República. https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm

Magramo, K. (2024, 17 de mayo). British engineering giant Arup revealed as $25 million deepfake scam victim. *CNN Business*. https://www.cnn.com/2024/05/16/tech/arup-deepfake-scam-loss-hong-kong-intl-hnk/index.html

Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. *ACM Computing Surveys, 54*(1), 1–41. https://doi.org/10.1145/3425780

Mitnick, K. D., & Simon, W. L. (2002). *The art of deception: Controlling the human element of security*. Wiley.

MITRE Corporation. (2026). *ATT&CK for enterprise* (Version 19.2). https://attack.mitre.org

National Institute of Standards and Technology. (2024). *The NIST Cybersecurity Framework (CSF) 2.0* (NIST CSWP 29). U.S. Department of Commerce. https://doi.org/10.6028/NIST.CSWP.29

Prakash, P. (2024, 17 de mayo). A deepfake ‘CFO’ tricked the British design firm behind the Sydney Opera House in $25 million scam. *Fortune*. https://fortune.com/europe/2024/05/17/arup-deepfake-fraud-scam-victim-hong-kong-25-million-cfo

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review, 9*(11), 40–53. https://doi.org/10.22215/timreview/1282

Yin, R. K. (2018). *Case study research and applications: Design and methods* (6th ed.). Sage.

Este artículo se publica en acceso abierto (Open Access) bajo licencia Creative Commons Attribution 4.0 Internacional (CC-BY), que permite uso, distribución y reproducción irrestrictos en cualquier medio, siempre que el autor original y la fuente sean acreditados.