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

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

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

Marcelo de Melo Marcon[1]

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

PUBLICADO: 10/2026

RESUMO

A convergência entre inteligência artificial generativa e engenharia social redefiniu o perfil de risco das organizações, deslocando o vetor de ataque do perímetro técnico para os processos de autorização e para os mecanismos humanos de verificação de identidade. Este artigo tem por objetivo inferir, a partir de evidências públicas, as lacunas de controle organizacional exploradas por esses ataques e propor controles compensatórios aplicáveis. Adota-se a estratégia de estudo de caso único holístico, de natureza qualitativa e base documental, tendo por unidade de análise o incidente ocorrido em janeiro de 2024 no escritório de Hong Kong de uma empresa multinacional britânica de engenharia, no qual um profissional da área financeira executou quinze transferências totalizando aproximadamente 25,6 milhões de dólares após participar de videoconferência em que todos os demais participantes eram sínteses audiovisuais geradas por inteligência artificial. As evidências foram obtidas exclusivamente em fontes públicas e submetidas a triangulação, com indicação das fontes que sustentam cada fato. A análise mapeia a cadeia de ataque contra as técnicas do MITRE ATT&CK, na versão 19.2, e confronta o incidente com a ISO/IEC 27001:2022 e com o NIST Cybersecurity Framework 2.0. Não foi reportado comprometimento tecnológico dos sistemas, e as evidências analisadas apontam para fragilidades de natureza procedimental, entre as quais a aceitação da evidência sensorial como prova de identidade. Propõe-se reposicionar a videoconferência como canal não autenticado.

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

ABSTRACT

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 ATT&CK techniques, version 19.2, and confronts the incident with ISO/IEC 27001:2022 and the NIST

[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 artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

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

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.

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

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 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.

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

INTRODUÇÃO

A doutrina de segurança da informação organizou-se, por décadas, em torno da defesa do perímetro tecnológico, arquitetura que pressupõe que o ataque deixe rastro técnico. Quando o adversário não invade sistema algum e obtém o mesmo resultado pela manipulação de um processo autorizativo legítimo, os controles perimetrais tornam-se irrelevantes por construção. A engenharia social sempre representou essa exceção. O que se alterou a partir de 2023 não foi a lógica do ataque, mas sua escala de credibilidade: modelos generativos de síntese audiovisual tornaram-se acessíveis, baratos e capazes de operar em tempo real. A evidência sensorial — ver o rosto de um colega e ouvir sua voz — deixou de constituir prova de identidade, embora continue a operar como tal na maior parte dos processos organizacionais.

Os dados agregados confirmam a materialidade do fenômeno. O Internet Crime Complaint Center do Federal Bureau of Investigation registrou, em 2025, mais de um milhão de reclamações e perdas superiores a 20,8 bilhões de dólares, das quais cerca de 3,0 bilhões atribuídos ao comprometimento de e-mail corporativo; no mesmo relatório, a inteligência artificial foi incorporada pela primeira vez como descritor formal de modalidade criminosa, respondendo por 22.364 reclamações e cerca de 893 milhões de dólares (Federal Bureau of Investigation [FBI], 2026).

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

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

Projeções do setor financeiro estimam que fraudes habilitadas por IA generativa alcancem 40 bilhões de dólares anuais nos Estados Unidos até 2027, ante 12,3 bilhões em 2023 (Deloitte, 2024).

Nesse contexto, o incidente sofrido por uma empresa multinacional britânica de engenharia em janeiro de 2024 constitui caso paradigmático. Um profissional da área financeira do escritório de Hong Kong executou quinze transferências, totalizando cerca de 25,6 milhões de dólares norte-americanos, após participar de videoconferência na qual acreditava dialogar com o diretor financeiro global e outros colegas — todos representações sintéticas geradas a partir de material audiovisual publicamente disponível. Nenhum sistema foi comprometido e nenhum dado foi exfiltrado (Magramo, 2024; Prakash, 2024). O caso é relevante porque a ausência de comprometimento técnico isola a variável de interesse: não há malware a analisar nem vulnerabilidade a corrigir. Resta a arquitetura de autorização financeira e os pressupostos de verificação de identidade nela embutidos.

Delimita-se a seguinte questão de pesquisa: quais lacunas de controle organizacional são exploradas por ataques de engenharia social assistidos por inteligência artificial generativa e quais controles são capazes de mitigá-las? O objetivo geral é identificar tais lacunas a partir da análise sistemática de um caso real, confrontando-o com frameworks consolidados. Como objetivos específicos: (i) reconstruir a cadeia de ataque a partir de evidências públicas trianguladas; (ii) mapear as técnicas segundo o MITRE ATT&CK; (iii) identificar as lacunas à luz do Anexo A da ISO/IEC 27001:2022; e (iv) propor controles compensatórios organizados pelas funções do NIST Cybersecurity Framework 2.0.

A justificativa apoia-se em três razões. No conjunto de literatura examinado pelos autores, a produção sobre deepfakes concentra-se na base técnica de geração e detecção (Goodfellow et al., 2014; Mirsky & Lee, 2021) e em implicações jurídicas e político-eleitorais (Westerlund, 2019; Chesney & Citron, 2019), sendo menos frequente o exame do fenômeno como problema de controle organizacional — delimitação que se refere ao corpus consultado e não a uma revisão sistemática da literatura. Em segundo lugar, a reconstrução documentada de um incidente de alto impacto oferece insumo direto para revisão de políticas de autorização financeira. Em terceiro, a erosão da confiança na evidência audiovisual afeta os mecanismos de verificação sobre os quais se assentam relações contratuais e institucionais. O artigo estrutura-se em cinco seções.

1 - REFERENCIAL TEÓRICO

1.1 ENGENHARIA SOCIAL E OS LIMITES DA CONSCIENTIZAÇÃO

A engenharia social designa as técnicas destinadas a induzir pessoas a executar ações ou revelar informações que comprometem a segurança, mediante manipulação da confiança e não da tecnologia. Mitnick e Simon (2002) demonstraram que o elemento humano é o vetor de menor resistência em qualquer arquitetura de segurança, por operar segundo heurísticas sociais que não

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

podem ser desativadas sem inviabilizar a cooperação organizacional. Hadnagy (2018) sistematiza o ataque em fases — coleta, pretexto, vínculo, exploração e execução —, ressaltando que a qualidade do pretexto depende do reconhecimento prévio.

O fundamento psicológico dessa eficácia foi mapeado por Cialdini (2009) em seis princípios de influência, dos quais três operam com particular intensidade em contextos corporativos: autoridade, escassez e prova social. A conscientização isolada, porém, apresenta eficácia limitada: treinamentos baseados em reconhecimento de indícios — erros gramaticais, domínios suspeitos, remetentes anômalos — pressupõem que o ataque contenha imperfeições detectáveis. Quando a síntese generativa elimina esses indícios, o modelo de defesa fundado na percepção individual perde sustentação, restando apenas os controles de processo, o que desloca o problema do domínio da conscientização para o da governança.

1.2 DEEPFAKES E INTELIGÊNCIA ARTIFICIAL GENERATIVA

O termo deepfake designa conteúdo audiovisual sintético produzido por arquiteturas de aprendizado profundo, tipicamente com o propósito de substituir ou reenactar a identidade de um indivíduo. A base técnica remonta às redes generativas adversárias propostas por Goodfellow et al. (2014), nas quais duas redes neurais competem — uma gerando amostras, outra discriminando entre sintéticas e reais —, elevando progressivamente a fidelidade da síntese. Mirsky e Lee (2021) alertam que os métodos de detecção exibem degradação significativa de desempenho quando aplicados a dados fora da distribuição de treinamento; a corrida entre geração e detecção favorece estruturalmente o gerador, pois qualquer detector publicado torna-se alvo de otimização adversária. Westerlund (2019) argumenta que o impacto do deepfake decorre menos de sua qualidade absoluta do que do colapso da barreira de entrada, e Chesney e Citron (2019) formulam o mesmo fenômeno em termos jurídicos, cunhando a noção de dividendo do mentiroso: a difusão de conteúdo sintético degrada a autoridade probatória de todo registro audiovisual, inclusive dos autênticos. Importa reter três atributos: a síntese em tempo real tornou-se viável em plataformas convencionais; o material de treinamento é obtido em fontes abertas; e a detecção automatizada não constitui, hoje, controle confiável em ambiente operacional.

1.3 FRAUDE DO CEO E REFERENCIAIS DE CONTROLE

O comprometimento de e-mail corporativo consiste na indução de transferência financeira mediante personificação de figura de autoridade interna ou de parceiro comercial. Segundo o FBI (2026), a categoria respondeu por 3,046 bilhões de dólares em perdas reportadas em 2025, em 24.768 reclamações, sendo a segunda maior por valor; o relatório anterior registrara 2,77 bilhões em 21.442 reclamações (FBI, 2025). A estrutura canônica compreende reconhecimento da hierarquia e dos fluxos de pagamento, personificação por canal assíncrono, pretexto de

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

confidencialidade e urgência, e direcionamento a contas do adversário. A defesa consagrada apoia-se em verificação fora de banda e em segregação de funções. A inovação do caso analisado está na extensão do ataque a canal síncrono e audiovisual: quando o canal considerado mais confiável passa a ser falsificável, o próprio controle compensatório é neutralizado, e não apenas contornado.

A análise recorre a três instrumentos consolidados. O MITRE ATT&CK é base de conhecimento de táticas e técnicas adversárias observadas empiricamente; a partir de 2023 passou a contemplar técnicas de personificação e de furto financeiro, tornando-se aplicável a incidentes sem comprometimento de sistemas. Na versão 19.2, consultada em outubro de 2026 e adotada neste trabalho, essas técnicas correspondem a T1684.001 (Impersonation), subtécnica de T1684 (Social Engineering), e a T1657 (Financial Theft) (MITRE Corporation, 2026). A ISO/IEC 27001:2022 estabelece requisitos para sistemas de gestão de segurança da informação (International Organization for Standardization [ISO], 2022). E o NIST Cybersecurity Framework 2.0 organiza resultados em seis funções, sendo relevante a introdução da função Governar, que explicita que decisões de risco e políticas antecedem os controles técnicos (National Institute of Standards and Technology [NIST], 2024).

2 - METODOLOGIA

2.1 DELINEAMENTO E UNIDADE DE ANÁLISE

A pesquisa é qualitativa quanto à abordagem, descritiva e explicativa quanto aos objetivos, e documental quanto aos procedimentos. Adota-se a estratégia de estudo de caso único holístico, conforme Yin (2018), justificada por dois critérios: o caráter revelador do fenômeno, por tratar-se de um dos primeiros incidentes de fraude mediada por síntese audiovisual em tempo real documentado publicamente e com atribuição confirmada pela vítima; e a magnitude do impacto, que assegura densidade documental incomum. O mapeamento de técnicas adota a versão 19.2 do MITRE ATT&CK for Enterprise, consultada em outubro de 2026.

A unidade de análise é o incidente ocorrido em janeiro de 2024 no escritório de Hong Kong da organização, delimitado entre a preparação do ataque e a divulgação pública da identidade da vítima, em maio de 2024. Ficam fora do escopo os desdobramentos investigativos posteriores.

2.2 FONTES DE EVIDÊNCIA E TRIANGULAÇÃO

As evidências foram obtidas exclusivamente em fontes públicas, o que dispensa submissão a Comitê de Ética em Pesquisa, uma vez que não houve coleta junto a seres humanos nem acesso a informações não divulgadas. Consultaram-se três categorias: reportagens de veículos jornalísticos de referência, nas quais se encontram reproduzidos o comunicado da organização vitimada e as informações divulgadas pela polícia de Hong Kong; relatórios de autoridades e organismos setoriais, empregados para caracterização do contexto e não do incidente; e literatura revisada por pares.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Registre-se que o comunicado corporativo e as declarações policiais não foram obtidos em documentos autônomos, mas nas reportagens que os reproduzem, razão pela qual são citados por intermédio delas.

Aplicou-se o critério de triangulação de Yin (2018): só foram incorporados os elementos factuais confirmados por, no mínimo, duas fontes independentes. Consideraram-se independentes as fontes de propriedade editorial distinta, com autoria própria e apuração não derivada de um mesmo despacho de agência — condição satisfeita por Magramo (2024), da CNN, e Prakash (2024), da Fortune. A coluna de fontes da Tabela 1 indica, por evento, quais documentos o sustentam. Informações de fornecedores de soluções de segurança não foram usadas como base factual. Por decisão dos autores, a organização vitimada não é nominalmente identificada no corpo deste artigo, embora os documentos jornalísticos a identifiquem e seus títulos sejam reproduzidos nas referências conforme publicados, pois a alteração de título constitui erro de citação.

2.3 LIMITAÇÕES DO ESTUDO

Não houve acesso a documentação interna, registros de auditoria ou relatórios forenses. Em consequência, as lacunas apresentadas na seção 3.3 constituem inferências fundamentadas na descrição pública do incidente, e não constatações de auditoria: as fontes disponíveis não permitem distinguir entre a inexistência de um controle, sua existência com contorno, a concessão de exceção operacional ou outra circunstância não documentada. Essa ressalva vale para todo o texto, e a terminologia de inferência é mantida nas tabelas e nas considerações finais. Registram-se outras três limitações: o estudo de caso único permite generalização analítica, mas não estatística (Yin, 2018); há possibilidade de viés de sobrevivência, pois tornam-se públicos os casos de maior magnitude; e a capacidade de síntese audiovisual evoluiu desde a ocorrência, de modo que as recomendações devem ser lidas como orientações de processo.

2.4 DECLARAÇÃO DO USO DE IA

Houve uso de IA: Os autores declaram que utilizaram o assistente Claude, da Anthropic, modelo Claude Opus 5, acessado entre setembro e outubro de 2026, nas etapas de levantamento e triagem de fontes, definição do delineamento, redação do manuscrito, reconstrução da cadeia de ataque, elaboração do mapeamento contra o MITRE ATT&CK, identificação das lacunas de controle, formulação dos controles compensatórios e formatação do documento. Registre-se, para precisão da declaração, que essa participação não se limitou a assistência textual: abrangeu a organização analítica do material e a proposição inicial do delineamento, do mapeamento, das lacunas e dos controles, cabendo aos autores a decisão sobre cada um desses elementos, sua verificação e sua validação. O uso afetou a totalidade do texto, incluídas as Tabelas 1 a 4 e as referências, e abrange a revisão feita em resposta ao parecer. São de responsabilidade exclusiva

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

dos autores a definição do tema e do recorte, a escolha do caso, as decisões sobre o delineamento, os critérios de triangulação, a decisão de não identificar nominalmente a organização e a validação final. Todas as informações sobre o incidente provêm de reportagens jornalísticas, nas quais se encontram reproduzidos o comunicado da organização e as declarações policiais; não foram fornecidos ao sistema dados internos, confidenciais ou obtidos por via não pública, e o profissional envolvido não é identificado nem caracterizado, razão pela qual não houve tratamento de dados sujeitos à Lei nº 13.709/2018. Os prompts e registros metodologicamente relevantes estão arquivados pelos autores e disponíveis à editoria mediante solicitação. As saídas foram revisadas mediante conferência de cada referência contra a fonte primária, dos identificadores digitais de objeto, dos códigos do MITRE ATT&CK contra a versão 19.2 da taxonomia e dos identificadores de controle da ISO/IEC 27001:2022. Essa conferência, refeita nesta revisão, levou à correção de dois enquadramentos do mapeamento, conforme a nota da Tabela 2. Nenhum dado, citação ou conclusão foi aceito sem verificação humana. Os autores assumem responsabilidade integral pelo conteúdo final.

3 - RESULTADOS E DISCUSSÃO

3.1 O CASO E A RECONSTRUÇÃO DA CADEIA DE ATAQUE

A organização é empresa multinacional britânica de consultoria em engenharia, sediada em Londres, que à época declarava operar com cerca de dezoito mil e quinhentos profissionais em trinta e quatro escritórios (Magramo, 2024). O perfil é relevante por duas razões: a dispersão geográfica reduz a familiaridade interpessoal entre matriz e unidades, tornando a comunicação mediada por tecnologia a norma; e a alta exposição pública dos executivos disponibiliza material audiovisual em fontes abertas. A Tabela 1 consolida a sequência de eventos.

Tabela 1

Cronologia do incidente

Momento	Evento	Fontes
Período prévio	Coleta de material audiovisual público de executivos, obtido em gravações de eventos, entrevistas e reuniões corporativas divulgadas, e utilizado para treinamento dos modelos de síntese.	Magramo (2024); Prakash (2024)
Janeiro de 2024	Recepção, por profissional da área financeira do escritório de Hong Kong, de mensagem eletrônica atribuída ao diretor financeiro sediado no Reino Unido, solicitando transação confidencial. O profissional manifesta suspeita e identifica indícios compatíveis com phishing.	Magramo (2024); Prakash (2024)
Janeiro de 2024	Realização de videoconferência com participação aparente do diretor financeiro e de outros colegas reconhecíveis, todos sinteticamente gerados. A solicitação é reiterada e confirmada no canal síncrono.	Magramo (2024); Prakash (2024)
Janeiro de 2024	Execução de quinze transferências, totalizando cerca de duzentos milhões de dólares de Hong Kong, para cinco contas distintas.	Magramo (2024); Prakash (2024)
Fevereiro de 2024	Descoberta da fraude em contato posterior do profissional com a sede, comunicação à polícia de Hong Kong e divulgação do caso pela polícia, sem identificação nominal da vítima.	Magramo (2024); Prakash (2024)

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

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

Momento	Evento	Fontes
Maio de 2024	Confirmação pública pela organização, com esclarecimento de que nenhum sistema interno foi comprometido e nenhum dado foi afetado.	Comunicado da organização reproduzido em Magramo (2024) e em Prakash (2024)

Nota. Elaborado pelos autores. A coluna de fontes indica os documentos que sustentam cada evento; todos os eventos relacionados contam com duas fontes de propriedade editorial distinta, conforme o critério de independência definido na subseção 2.2. Os dados de contexto de Deloitte (2024) e FBI (2026) não sustentam fatos do incidente e por isso não figuram nesta tabela.

A reconstrução permite identificar cinco fases. O reconhecimento consistiu na identificação da estrutura hierárquica, do fluxo de autorização e do profissional competente, e na coleta do material audiovisual — fase passiva, sobre fontes abertas, indetectável pela organização alvo. O desenvolvimento de recursos envolveu cinco contas receptoras, o que fraciona os valores, reduz a probabilidade de acionamento de limiares antifraude e distribui o risco de bloqueio. O contato inicial ocorreu por canal assíncrono e falhou: o profissional identificou a anomalia e não executou a solicitação, de modo que o controle humano funcionou conforme o esperado.

A superação da resistência é o elemento distintivo do caso. Diante da suspeita manifestada, o adversário não abandonou o alvo nem intensificou a pressão pelo mesmo canal, mas migrou para canal de maior credibilidade percebida. A videoconferência com múltiplos participantes reconhecíveis mobilizou três princípios de influência: autoridade, prova social e escassez. É precisamente o mecanismo de verificação adotado pelo profissional que foi convertido em instrumento de persuasão. A descoberta ocorreu por iniciativa do próprio profissional, e não por controle automatizado.

3.2 MAPEAMENTO SEGUNDO O MITRE ATT&CK

Tabela 2

Mapeamento da cadeia de ataque segundo o MITRE ATT&CK

Fase	Técnica	Evidência no caso
Reconhecimento	T1589 — Gather Victim Identity Information	Identificação de executivos e do profissional responsável pela execução de pagamentos.
Reconhecimento	T1593 — Search Open Websites/Domains	Obtenção, em sítios corporativos e redes sociais, do material audiovisual público utilizado no treinamento dos modelos de síntese.
Acesso inicial	T1566 — Phishing	Mensagem eletrônica atribuída ao diretor financeiro solicitando transação confidencial.
Stealth	T1684.001 — Impersonation	Personificação audiovisual em tempo real do diretor financeiro e de colegas em videoconferência.
Impacto	T1657 — Financial Theft	Execução de quinze transferências para cinco contas bancárias previamente constituídas e controladas pelo adversário.

Nota. Elaborado pelos autores com base em MITRE Corporation (2026), versão 19.2 da taxonomia, consultada em outubro de 2026, e nas evidências do caso. Dois enquadramentos da versão anterior deste mapeamento foram corrigidos na revisão. A personificação constava como T1656; na versão 19.2 essa técnica integra T1684 (Social Engineering) como subtécnica T1684.001, sob a tática Stealth, antes denominada Evasão de Defesas. E a constituição das contas bancárias receptoras constava como T1585 (Establish Accounts), enquadramento retirado por não corresponder à definição da técnica, restrita a contas em serviços online destinadas à construção de persona; a evidência foi realocada para T1657, cuja descrição contempla o envio de recursos a contas controladas pelo adversário.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

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

Três observações decorrem do mapeamento. A primeira é a ausência de técnicas das táticas de execução, persistência, movimentação lateral e exfiltração: a cadeia transita do acesso inicial à personificação e ao impacto, sem interação com sistemas da organização. A segunda é que a técnica decisiva, T1684.001, não possui contramedida técnica convencional — não há artefato a detectar no endpoint nem tráfego anômalo na rede —, e a mitigação prevista é processual. A terceira é que a constituição prévia de contas bancárias receptoras, embora documentada no caso e relevante para a logística do ataque, não encontra técnica correspondente na matriz Enterprise, construída para descrever a intrusão em sistemas e não a cadeia financeira que monetiza a fraude — limite da taxonomia que o próprio enquadramento deste caso torna visível.

3.3 LACUNAS DE CONTROLE À LUZ DA ISO/IEC 27001:2022

Tabela 3

Lacunas de controle inferidas segundo o Anexo A da ISO/IEC 27001:2022

Controle	Resultado esperado	Lacuna inferida
5.3 — Segregação de funções	Impossibilidade de que uma única pessoa inicie e conclua transação de alto valor.	A execução de quinze transferências de valor agregado elevado por um único profissional é compatível com a hipótese de ausência de segregação vinculante ou de sua suspensão mediante invocação de confidencialidade. As fontes públicas não permitem determinar o fluxo interno de aprovação nem distinguir entre inexistência do controle, seu contorno e eventual exceção autorizada.
5.14 — Transferência de informações	Definição de canais autorizados e de regras de verificação para comunicações sensíveis.	A aceitação da videoconferência como confirmação suficiente sugere ausência de exigência de canal independente previamente estabelecido.
5.15 e 5.17 — Controle de acesso e autenticação	Verificação de identidade baseada em fatores não replicáveis.	A verificação apoiou-se em reconhecimento sensorial de face e voz, atributos publicamente disponíveis e sinteticamente reproduzíveis.
8.16 — Atividades de monitoramento	Detecção de padrões transacionais anômalos em tempo hábil.	A detecção ocorreu por iniciativa humana posterior, sem registro público de alerta por fracionamento ou destino atípico.

Nota. Elaborado pelos autores com base em International Organization for Standardization (2022) e nas evidências públicas do caso. As lacunas são inferências dos autores a partir do desfecho documentado, e não constatações de auditoria.

Somam-se às lacunas da tabela outras duas, igualmente inferidas: o controle 5.7, de inteligência de ameaças, aparentemente não contemplava a síntese audiovisual em tempo real como vetor previsto no processo de pagamento; e o controle 6.3, de conscientização, orientava-se à identificação de indícios em canais assíncronos, sem cobertura de personificação síncrona. A lacuna mais consequente é a relativa à verificação de identidade: os demais controles pressupõem mecanismo confiável para estabelecer quem está do outro lado do canal, e quando esse mecanismo é a percepção sensorial, tornada reproduzível, a cadeia de controles dependentes fica comprometida. A expressão lacuna inferida é usada em sentido estrito: não se afirma que a organização não possuísse políticas formais, mas que o desfecho documentado é mais compatível com a hipótese de falha de eficácia desses controles do que com a de sua operação plena.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

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

3.4 A OBSOLESCÊNCIA DA EVIDÊNCIA SENSORIAL COMO PROVA DE IDENTIDADE

A contribuição analítica central refere-se ao pressuposto implícito que sustenta os processos autorizativos organizacionais. A proposição falei com a pessoa operou historicamente como padrão-ouro de confirmação, sem jamais ser explicitada. O caso oferece evidência de que esse pressuposto não se sustenta diante da síntese audiovisual em tempo real, e de forma aguda, porque o profissional não agiu com negligência: identificou o indício no canal assíncrono, manifestou suspeita e recorreu a mecanismo de verificação. O que falhou não foi o julgamento individual, mas a validade do instrumento disponível — distinção relevante, pois atribuir o incidente a erro humano conduz a respostas centradas em treinamento, quando a correção necessária é estrutural.

Formula-se, a partir dessa análise, a proposição que sintetiza o resultado do estudo: em ambientes nos quais a síntese audiovisual em tempo real é tecnicamente acessível, a videoconferência deveria ser classificada como canal não autenticado, equiparando-se, para fins de controle, à mensagem eletrônica e à chamada telefônica. Trata-se de proposição analítica e de recomendação de governança derivadas deste caso e da literatura, e não de regra cuja validade universal um estudo de caso único possa demonstrar. Dela decorre que a confirmação por videoconferência não deveria ser aceita como satisfação do requisito de verificação fora de banda em autorização financeira. A proposição contraria a orientação, comum em materiais de conscientização, de identificar deepfakes solicitando gestos ao interlocutor — recomendação que transfere ao indivíduo o ônus de uma perícia que não lhe compete e cuja validade é decrescente, por constituir alvo explícito de aperfeiçoamento dos geradores (Mirsky & Lee, 2021).

3.5 CONTROLES COMPENSATÓRIOS PROPOSTOS

Tabela 4

Controles compensatórios propostos por função do NIST CSF 2.0

Função	Controle proposto	Efeito esperado
Governar	Classificar formalmente a videoconferência como canal não autenticado em política de autorização financeira.	Explicita o pressuposto de confiabilidade e torna a decisão de risco auditável.
Governar	Estabelecer que solicitações de sigilo interno não suspendem controles de autorização.	Reduz a eficácia do pretexto de confidencialidade, elemento recorrente da modalidade.
Proteger	Exigir verificação fora de banda por canal previamente registrado, iniciada pelo verificador, para transações acima de limiar definido.	Dificulta que o adversário controle o canal de confirmação.
Proteger	Instituir segregação de funções vinculante, com aprovação por dois responsáveis independentes.	Eleva o custo do ataque, por exigir o comprometimento simultâneo de dois decisores.
Proteger	Adotar palavra-código ou desafio compartilhado, definido previamente e por canal distinto, para transações excepcionais.	Introduz fator de verificação não derivável de material público.
Detectar	Implementar regras de correlação para fracionamento de valores, destinos novos e desvio de padrão histórico.	Pode ampliar a probabilidade de interrupção durante a janela de execução.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

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

Função	Controle proposto	Efeito esperado
Responder	Acionar o canal de recuperação bancária no menor prazo possível, segundo meta definida na política da organização, e instituir política de não punição para comunicação tempestiva de suspeita ou erro.	Pode contribuir para reduzir o risco de dispersão internacional dos valores e a latência entre suspeita e relato.

Nota. Elaborado pelos autores com base em National Institute of Standards and Technology (2024), International Organization for Standardization (2022) e na análise do caso. Os efeitos indicados são proposições derivadas da análise; os controles não foram implementados nem avaliados empiricamente neste estudo.

Nenhum dos controles propostos depende de detecção automatizada de conteúdo sintético: os detectores disponíveis degradam fora do domínio de treinamento e são alvo de otimização adversária, de modo que fundamentar a defesa em detecção transferiria a segurança do processo para uma corrida tecnológica desfavorável ao defensor.

3.6 DISCUSSÃO DOS RESULTADOS

Os resultados convergem com a literatura quanto ao papel dos princípios de autoridade e escassez, mas acrescentam elemento não suficientemente tratado: a mobilização da prova social pela presença simultânea de múltiplos interlocutores sintéticos — amplificação qualitativa, e não quantitativa. Um interlocutor falsificado exige do alvo a suspensão de uma verificação; vários interlocutores convergentes exigem a suspensão da hipótese de que tal coordenação seja factível, hipótese que o custo marginal quase nulo da síntese adicional tornou obsoleta. Os achados dialogam ainda com Chesney e Citron (2019) sobre a degradação da autoridade probatória do registro audiovisual: se a videoconferência não prova identidade, práticas correntes como aprovações remotas, entrevistas de contratação e recuperação de credenciais junto ao suporte operam sobre fundamento comprometido.

Do ponto de vista da gestão de riscos, a resposta adequada não é necessariamente tecnológica: a assimetria entre geração e detecção favorece o adversário, ao passo que controles de processo permanecem eficazes independentemente da qualidade da falsificação, porque não dependem de distinguir o verdadeiro do falso, mas de exigir do adversário o controle de canais que ele não controla. Sugere-se, por fim, revisão dos programas de conscientização: os orientados ao reconhecimento de indícios tendem à obsolescência, enquanto os orientados à disciplina de procedimento mostram-se mais robustos, por exigirem que o profissional aja contra a própria percepção.

4 - CONSIDERAÇÕES

Este trabalho examinou, por estudo de caso único holístico de base documental, as lacunas de controle exploradas por ataque de engenharia social assistido por inteligência artificial generativa, tendo por objeto o incidente de janeiro de 2024 no escritório de Hong Kong de uma empresa multinacional britânica de engenharia.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

A cadeia de ataque foi reconstruída em cinco fases a partir de evidências públicas trianguladas, evidenciando-se que a fase decisiva não foi o contato inicial, mas a superação da resistência legitimamente manifestada pelo profissional. O mapeamento segundo o MITRE ATT&CK revelou cadeia atipicamente curta, sem qualquer interação com sistemas da organização, o que indica a inadequação dos controles perimetrais para esta classe de ameaça. Foram inferidas seis lacunas de controle a partir das evidências públicas disponíveis, à luz da ISO/IEC 27001:2022, das quais a mais consequente refere-se à verificação de identidade, e propôs-se conjunto de controles compensatórios organizados pelas funções do NIST Cybersecurity Framework 2.0, todos independentes de detecção automatizada de conteúdo sintético.

A contribuição teórica consiste na proposição de que a videoconferência deveria ser reclassificada como canal não autenticado, com a consequência de que a confirmação por vídeo não satisfaz o requisito de verificação fora de banda. A contribuição prática reside no conjunto de controles proposto, predominantemente processual e de governança, cuja adoção depende sobretudo de decisão quanto à supressão de exceções procedimentais; o estudo não avaliou custos, infraestrutura ou impacto operacional de sua implementação. Reconhecem-se as limitações já explicitadas, em especial a ausência de acesso a documentação interna, que confere às lacunas o estatuto de inferência.

Como agenda futura, sugerem-se quatro direções: estudo de casos múltiplos comparando incidentes de personificação sintética em diferentes setores; survey sobre o grau de adoção de verificação fora de banda em organizações brasileiras; estudo experimental da eficácia comparada entre programas de conscientização orientados a indícios e a disciplina de procedimento; e análise das implicações jurídicas da personificação sintética no direito brasileiro, quanto à responsabilidade do preposto que executa ordem falsificada de modo indetectável e ao tratamento dos dados biométricos usados no treinamento dos modelos à luz da LGPD.

REFERÊNCIAS

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 artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

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 maio). 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 maio). 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 artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.