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

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

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

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

PUBLISHED: 10/2026

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 such attacks and proposes applicable compensating controls. A holistic single-case study of a qualitative, documentary nature 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 Cybersecurity Framework 2.0. No technological compromise of the organization's systems was reported, and the evidence examined points to weaknesses of a procedural nature, 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.

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 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

[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).

This article is published in Open Access under the Creative Commons Attribution 4.0 International (CC-BY) license, which allows unrestricted use, distribution and reproduction in any medium, provided the original author and source are credited.

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.

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.

INTRODUCTION

For decades, information security doctrine was organized around the defense of the technological perimeter, an architecture that presumes an attack leaves a technical trace. When the adversary breaks into no system at all and obtains the same result by manipulating a legitimate authorization process, perimeter controls become irrelevant by construction. Social engineering has always represented that exception. What changed from 2023 onward was not the logic of the attack but its scale of credibility: generative models for audiovisual synthesis became accessible, inexpensive, and capable of operating in real time. Sensory evidence — seeing a colleague's face and hearing their voice — ceased to constitute proof of identity, even though it continues to operate as such in most organizational processes.

Aggregate data confirm the materiality of the phenomenon. The Internet Crime Complaint Center of the Federal Bureau of Investigation recorded, in 2025, more than one million complaints and losses exceeding USD 20.8 billion, of which roughly USD 3.0 billion were attributed to business email compromise; in the same report, artificial intelligence was incorporated for the first time as a formal descriptor of criminal modality, accounting for 22,364 complaints and approximately USD 893 million (Federal Bureau of Investigation [FBI], 2026). Financial-sector projections estimate that fraud enabled by generative AI will reach USD 40 billion annually in the United States by 2027, up from USD 12.3 billion in 2023 (Deloitte, 2024).

This article is published in Open Access under the Creative Commons Attribution 4.0 International (CC-BY) license, which allows unrestricted use, distribution and reproduction in any medium, provided the original author and source are credited.

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

Against that backdrop, the incident suffered by a British multinational engineering firm in January 2024 constitutes a paradigmatic case. A finance professional in the Hong Kong office executed fifteen transfers totaling approximately USD 25.6 million after attending a video conference in which they believed they were speaking with the global chief financial officer and other colleagues — all synthetic representations generated from publicly available audiovisual material. No system was compromised and no data were exfiltrated (Magramo, 2024; Prakash, 2024). The case is relevant because the absence of technical compromise isolates the variable of interest: there is no malware to analyze and no vulnerability to patch. What remains is the financial authorization architecture and the identity-verification assumptions embedded in it.

The following research question is posed: which organizational control gaps are exploited by social engineering attacks assisted by generative artificial intelligence, and which controls are capable of mitigating them? The general objective is to identify such gaps through the systematic analysis of a real case, confronting it with established frameworks. The specific objectives are: (i) to reconstruct the attack chain from triangulated public evidence; (ii) to map the techniques according to MITRE ATT&CK; (iii) to identify the gaps in light of Annex A of ISO/IEC 27001:2022; and (iv) to propose compensating controls organized by the functions of the NIST Cybersecurity Framework 2.0.

The rationale rests on three grounds. Within the body of literature examined by the authors, scholarship on deepfakes concentrates on the technical basis of generation and detection (Goodfellow et al., 2014; Mirsky & Lee, 2021) and on legal and electoral-political implications (Westerlund, 2019; Chesney & Citron, 2019), with examination of the phenomenon as a problem of organizational control being less frequent — a delimitation that refers to the corpus consulted and not to a systematic review of the literature. Second, the documented reconstruction of a high-impact incident offers direct input for revising financial authorization policies. Third, the erosion of trust in audiovisual evidence affects the verification mechanisms on which contractual and institutional relations rest. The article is organized into five sections.

1 - THEORETICAL FRAMEWORK

1.1 SOCIAL ENGINEERING AND THE LIMITS OF AWARENESS

Social engineering designates the techniques intended to induce people to perform actions or disclose information that compromises security, through the manipulation of trust rather than of technology. Mitnick and Simon (2002) showed that the human element is the path of least resistance in any security architecture, because it operates according to social heuristics that cannot be switched off without making organizational cooperation unworkable. Hadnagy (2018) systematizes the attack into phases — collection, pretext, rapport, exploitation, and execution — stressing that the quality of the pretext depends on prior reconnaissance.

This article is published in Open Access under the Creative Commons Attribution 4.0 International (CC-BY) license, which allows unrestricted use, distribution and reproduction in any medium, provided the original author and source are credited.

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

The psychological basis of that effectiveness was mapped by Cialdini (2009) into six principles of influence, three of which operate with particular intensity in corporate settings: authority, scarcity, and social proof. Awareness alone, however, has limited efficacy: training based on recognizing cues — grammatical errors, suspicious domains, anomalous senders — presumes that the attack contains detectable imperfections. When generative synthesis eliminates those cues, a defense model founded on individual perception loses its footing, leaving only process controls, which shifts the problem from the domain of awareness to that of governance.

1.2 DEEPFAKES AND GENERATIVE ARTIFICIAL INTELLIGENCE

The term deepfake designates synthetic audiovisual content produced by deep learning architectures, typically for the purpose of replacing or reenacting an individual's identity. The technical basis traces back to the generative adversarial networks proposed by Goodfellow et al. (2014), in which two neural networks compete — one generating samples, the other discriminating between synthetic and real — progressively raising the fidelity of the synthesis. Mirsky and Lee (2021) warn that detection methods exhibit significant performance degradation when applied to data outside the training distribution; the race between generation and detection structurally favors the generator, since any published detector becomes a target of adversarial optimization. Westerlund (2019) argues that the impact of the deepfake stems less from its absolute quality than from the collapse of the barrier to entry, and Chesney and Citron (2019) frame the same phenomenon in legal terms, coining the notion of the liar's dividend: the spread of synthetic content degrades the evidentiary authority of every audiovisual record, including authentic ones. Three attributes bear retaining: real-time synthesis has become feasible on conventional platforms; training material is obtained from open sources; and automated detection does not today constitute a reliable control in an operational environment.

1.3 CEO FRAUD AND CONTROL FRAMEWORKS

Business email compromise consists in inducing a financial transfer through the impersonation of an internal figure of authority or of a commercial partner. According to the FBI (2026), the category accounted for USD 3.046 billion in reported losses in 2025, across 24,768 complaints, the second largest by value; the previous report recorded USD 2.77 billion across 21,442 complaints (FBI, 2025). The canonical structure comprises reconnaissance of the hierarchy and of payment flows, impersonation over an asynchronous channel, a pretext of confidentiality and urgency, and routing to adversary-controlled accounts. The established defense rests on out-of-band verification and segregation of duties. The novelty of the case analyzed lies in the extension of the attack to a synchronous, audiovisual channel: when the channel regarded as most trustworthy becomes falsifiable, the compensating control itself is neutralized, not merely circumvented.

This article is published in Open Access under the Creative Commons Attribution 4.0 International (CC-BY) license, which allows unrestricted use, distribution and reproduction in any medium, provided the original author and source are credited.

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

The analysis draws on three established instruments. MITRE ATT&CK is a knowledge base of empirically observed adversary tactics and techniques; from 2023 onward it came to encompass techniques of impersonation and financial theft, making the taxonomy applicable to incidents without system compromise. In version 19.2, consulted in October 2026 and adopted in this work, those techniques correspond to T1684.001 (Impersonation), a sub-technique of T1684 (Social Engineering), and to T1657 (Financial Theft) (MITRE Corporation, 2026). ISO/IEC 27001:2022 establishes requirements for information security management systems (International Organization for Standardization [ISO], 2022). And the NIST Cybersecurity Framework 2.0 organizes outcomes into six functions, the introduction of the Govern function being relevant here, as it makes explicit that risk decisions and policies precede technical controls (National Institute of Standards and Technology [NIST], 2024).

2 - METHODOLOGY

2.1 RESEARCH DESIGN AND UNIT OF ANALYSIS

The research is qualitative in approach, descriptive and explanatory in its objectives, and documentary in its procedures. The strategy of a holistic single-case study is adopted, following Yin (2018), justified by two criteria: the revelatory character of the phenomenon, as it is one of the first publicly documented incidents of fraud mediated by real-time audiovisual synthesis with attribution confirmed by the victim; and the magnitude of the impact, which ensures unusual documentary density. The mapping of techniques adopts version 19.2 of MITRE ATT&CK for Enterprise, consulted in October 2026.

The unit of analysis is the incident that occurred in January 2024 at the organization's Hong Kong office, bounded between the preparation of the attack and the public disclosure of the victim's identity in May 2024. Subsequent investigative developments fall outside the scope.

2.2 SOURCES OF EVIDENCE AND TRIANGULATION

Evidence was obtained exclusively from public sources, which dispenses with submission to a Research Ethics Committee, since there was no collection from human subjects and no access to undisclosed information. Three categories were consulted: reports from reference journalistic outlets, in which the victim organization's statement and the information released by the Hong Kong police are reproduced; reports from authorities and sector bodies, used to characterize the context and not the incident; and peer-reviewed literature. It should be recorded that the corporate statement and the police declarations were not obtained as standalone documents but in the news reports that reproduce them, and are therefore cited by way of those reports.

Yin's (2018) triangulation criterion was applied: only factual elements confirmed by at least two independent sources were incorporated. Sources were considered independent when under distinct editorial ownership, with their own authorship and reporting not derived from a single wire

This article is published in Open Access under the Creative Commons Attribution 4.0 International (CC-BY) license, which allows unrestricted use, distribution and reproduction in any medium, provided the original author and source are credited.

dispatch — a condition satisfied by Magramo (2024), of CNN, and Prakash (2024), of Fortune. The sources column in Table 1 indicates, for each event, which documents support it. Information from security solution vendors was not used as a factual basis. By decision of the authors, the victim organization is not identified by name in the body of this article, although the journalistic documents identify it and their titles are reproduced in the references as published, since altering a title constitutes a citation error.

2.3 LIMITATIONS OF THE STUDY

There was no access to internal documentation, audit records, or forensic reports. Consequently, the gaps presented in section 3.3 constitute inferences grounded in the public description of the incident, and not audit findings: the available sources do not permit distinguishing between the nonexistence of a control, its existence with circumvention, the granting of an operational exception, or some other undocumented circumstance. That caveat holds for the entire text, and the terminology of inference is maintained in the tables and in the concluding remarks. Three further limitations are recorded: the single-case study permits analytical but not statistical generalization (Yin, 2018); survivorship bias is possible, since the cases that become public are those of greatest magnitude; and audiovisual synthesis capability has evolved since the occurrence, so that the recommendations should be read as process guidance.

2.4 DECLARATION OF ARTIFICIAL INTELLIGENCE USE

AI was used: The authors declare that they used the assistant Claude, from Anthropic, model Claude Opus 5, accessed between September and October 2026, in the survey and screening of sources, the definition of the research design, the writing of the manuscript, the reconstruction of the attack chain, the preparation of the mapping against MITRE ATT&CK, the identification of control gaps, the formulation of compensating controls, and the formatting of the document. For the precision of this declaration, it is recorded that such participation was not limited to textual assistance: it also encompassed the analytical organization of the material and the initial proposition of the research design, the mapping, the gaps, and the controls, with the authors deciding on each of those elements and carrying out their verification and validation. The use affected the entirety of the text, including Tables 1 through 4 and the references, and extends to the revision made in response to the review report. The definition of the topic and its delimitation, the choice of case, the decisions on research design, the triangulation criteria, the decision not to identify the organization by name, and the final validation are the exclusive responsibility of the authors. All information about the incident comes from news reports, in which the organization's statement and the police declarations are reproduced; no internal, confidential, or non-publicly obtained data were supplied to the system, and the professional involved is neither identified nor characterized, which is why there was no processing of data subject to Brazilian Law No. 13,709/2018 (General Personal Data Protection Law). The

This article is published in Open Access under the Creative Commons Attribution 4.0 International (CC-BY) license, which allows unrestricted use, distribution and reproduction in any medium, provided the original author and source are credited.

prompts and methodologically relevant records are archived by the authors and available to the editorial office upon request. The outputs were reviewed by checking every reference against the primary source, the digital object identifiers, the MITRE ATT&CK codes against version 19.2 of the taxonomy, and the ISO/IEC 27001:2022 control identifiers. That check, repeated in this revision, led to the correction of two mappings, as recorded in the note to Table 2. No datum, citation, or conclusion was accepted without human verification. The authors assume full responsibility for the final content.

3 - RESULTS AND DISCUSSION

3.1 THE CASE AND THE RECONSTRUCTION OF THE ATTACK CHAIN

The organization is a British multinational engineering consultancy headquartered in London which at the time stated that it operated with about eighteen thousand five hundred professionals across thirty-four offices (Magramo, 2024). The profile is relevant for two reasons: geographic dispersion reduces interpersonal familiarity between headquarters and units, making technology-mediated communication the norm; and the high public exposure of executives makes audiovisual material available from open sources. Table 1 consolidates the sequence of events.

Table 1

Chronology of the incident

Moment	Event	Sources
Prior period	Collection of public audiovisual material of executives, obtained from recordings of events, interviews, and published corporate meetings, and used to train the synthesis models.	Magramo (2024); Prakash (2024)
January 2024	Receipt, by a finance professional in the Hong Kong office, of an electronic message attributed to the chief financial officer based in the United Kingdom, requesting a confidential transaction. The professional expresses suspicion and identifies cues consistent with phishing.	Magramo (2024); Prakash (2024)
January 2024	Holding of a video conference with the apparent participation of the chief financial officer and other recognizable colleagues, all synthetically generated. The request is reiterated and confirmed on the synchronous channel.	Magramo (2024); Prakash (2024)
January 2024	Execution of fifteen transfers, totaling about two hundred million Hong Kong dollars, to five distinct accounts.	Magramo (2024); Prakash (2024)
February 2024	Discovery of the fraud during the professional's subsequent contact with headquarters, report to the Hong Kong police, and disclosure of the case by the police, without naming the victim.	Magramo (2024); Prakash (2024)
May 2024	Public confirmation by the organization, clarifying that no internal system was compromised and no data were affected.	Organization's statement as reproduced in Magramo (2024) and Prakash (2024)

Note. Compiled by the authors. The sources column indicates the documents supporting each event; all events listed are supported by two sources under distinct editorial ownership, per the independence criterion defined in subsection 2.2. The contextual data from Deloitte (2024) and the FBI (2026) do not support facts of the incident and therefore do not appear in this table.

The reconstruction allows five phases to be identified. Reconnaissance consisted in identifying the hierarchical structure, the authorization flow, and the competent professional, and in

This article is published in Open Access under the Creative Commons Attribution 4.0 International (CC-BY) license, which allows unrestricted use, distribution and reproduction in any medium, provided the original author and source are credited.

collecting the audiovisual material — a passive phase, over open sources, undetectable by the target organization. Resource development involved five receiving accounts, which fragments the amounts, reduces the probability of triggering antifraud thresholds, and distributes the risk of blocking. Initial contact occurred over an asynchronous channel and failed: the professional identified the anomaly and did not execute the request, so that the human control worked as expected.

Overcoming that resistance is the distinctive element of the case. Faced with the suspicion expressed, the adversary neither abandoned the target nor intensified pressure through the same channel, but migrated to a channel of greater perceived credibility. The video conference with multiple recognizable participants mobilized three principles of influence: authority, social proof, and scarcity. It is precisely the verification mechanism adopted by the professional that was converted into an instrument of persuasion. Discovery came at the professional's own initiative, not through an automated control.

3.2 MAPPING ACCORDING TO MITRE ATT&CK

Table 2

Mapping of the attack chain according to MITRE ATT&CK

Phase	Technique	Evidence in the case
Reconnaissance	T1589 — Gather Victim Identity Information	Identification of executives and of the professional responsible for executing payments.
Reconnaissance	T1593 — Search Open Websites/Domains	Obtaining, from corporate sites and social networks, the public audiovisual material used to train the synthesis models.
Initial Access	T1566 — Phishing	Electronic message attributed to the chief financial officer requesting a confidential transaction.
Stealth	T1684.001 — Impersonation	Real-time audiovisual impersonation of the chief financial officer and of colleagues in a video conference.
Impact	T1657 — Financial Theft	Execution of fifteen transfers to five bank accounts previously established and controlled by the adversary.

Note. Compiled by the authors based on MITRE Corporation (2026), version 19.2 of the taxonomy, consulted in October 2026, and on the evidence of the case. Two mappings from the previous version of this analysis were corrected in the revision. Impersonation was listed as T1656; in version 19.2 that technique forms part of T1684 (Social Engineering) as sub-technique T1684.001, under the Stealth tactic, formerly named Defense Evasion. And the establishment of the receiving bank accounts was listed as T1585 (Establish Accounts), a mapping withdrawn because it does not correspond to the definition of the technique, which is restricted to accounts with online services intended for persona building; the evidence was reallocated to T1657, whose description covers the sending of funds to adversary-controlled accounts.

Three observations follow from the mapping. The first is the absence of techniques from the execution, persistence, lateral movement, and exfiltration tactics: the chain moves from initial access to impersonation and impact, without interaction with the organization's systems. The second is that the decisive technique, T1684.001, has no conventional technical countermeasure — there is no artifact to detect on the endpoint and no anomalous traffic on the network — and the prescribed mitigation is procedural. The third is that the prior establishment of receiving bank accounts, although documented in the case and relevant to the logistics of the attack, has no corresponding technique in the Enterprise matrix, which was built to describe intrusion into systems rather than the financial

This article is published in Open Access under the Creative Commons Attribution 4.0 International (CC-BY) license, which allows unrestricted use, distribution and reproduction in any medium, provided the original author and source are credited.

chain that monetizes the fraud — a limit of the taxonomy that the mapping of this very case makes visible.

3.3 CONTROL GAPS IN LIGHT OF ISO/IEC 27001:2022

Table 3

Control gaps inferred under Annex A of ISO/IEC 27001:2022

Control	Expected outcome	Inferred gap
5.3 — Segregation of duties	Impossibility of a single person initiating and completing a high-value transaction.	The execution of fifteen transfers of high aggregate value by a single professional is consistent with the hypothesis of an absence of binding segregation, or of its suspension upon invocation of confidentiality. Public sources do not permit determining the internal approval flow, nor distinguishing between nonexistence of the control, its circumvention, and a possible authorized exception.
5.14 — Transfer of information	Definition of authorized channels and verification rules for sensitive communications.	The acceptance of the video conference as sufficient confirmation suggests the absence of any requirement for a previously established independent channel.
5.15 and 5.17 — Access control and authentication	Identity verification based on non-replicable factors.	Verification relied on sensory recognition of face and voice, attributes that are publicly available and synthetically reproducible.
8.16 — Monitoring activities	Timely detection of anomalous transactional patterns.	Detection occurred through subsequent human initiative, with no public record of an alert for fragmentation or atypical destination.

Note. Compiled by the authors based on International Organization for Standardization (2022) and on public evidence of the case. The gaps are inferences drawn by the authors from the documented outcome, and not audit findings.

Two further gaps, equally inferred, may be added to those in the table: control 5.7, on threat intelligence, apparently did not contemplate real-time audiovisual synthesis as a vector foreseen in the payment process; and control 6.3, on awareness, was oriented toward identifying cues in asynchronous channels, without covering synchronous impersonation. The most consequential gap concerns identity verification: the remaining controls presuppose a reliable mechanism for establishing who is on the other side of the channel, and when that mechanism is sensory perception, now reproducible, the chain of dependent controls is compromised. The expression inferred gap is used in a strict sense: it is not asserted that the organization lacked formal policies, but that the documented outcome is more consistent with the hypothesis of a failure in the effectiveness of those controls than with that of their full operation.

3.4 THE OBSOLESCENCE OF SENSORY EVIDENCE AS PROOF OF IDENTITY

The central analytical contribution concerns the implicit assumption that sustains organizational authorization processes. The proposition I spoke with the person historically operated as the gold standard of confirmation, without ever being made explicit. The case offers evidence that this assumption does not hold in the face of real-time audiovisual synthesis, and acutely so, because the professional did not act negligently: they identified the cue on the asynchronous channel,

This article is published in Open Access under the Creative Commons Attribution 4.0 International (CC-BY) license, which allows unrestricted use, distribution and reproduction in any medium, provided the original author and source are credited.

expressed suspicion, and resorted to a verification mechanism. What failed was not individual judgment but the validity of the available instrument — a relevant distinction, since attributing the incident to human error leads to responses centered on training, when the necessary correction is structural.

From this analysis the proposition that synthesizes the study's result is formulated: in environments where real-time audiovisual synthesis is technically accessible, video conferencing should be classified as an unauthenticated channel, placed on a par, for control purposes, with electronic messaging and the telephone call. This is an analytical proposition and a governance recommendation derived from this case and from the literature, and not a rule whose universal validity a single-case study could demonstrate. It follows that confirmation by video conference should not be accepted as satisfying the out-of-band verification requirement in financial authorization. The proposition runs counter to the guidance, common in awareness materials, of identifying deepfakes by asking the interlocutor to perform gestures — a recommendation that transfers to the individual the burden of an expert assessment that is not theirs to make, and whose validity is declining, since it constitutes an explicit target for the improvement of generators (Mirsky & Lee, 2021).

3.5 PROPOSED COMPENSATING CONTROLS

Table 4

Compensating controls proposed by NIST CSF 2.0 function

Function	Proposed control	Expected effect
Govern	Formally classify video conferencing as an unauthenticated channel in the financial authorization policy.	Makes the reliability assumption explicit and renders the risk decision auditable.
Govern	Establish that requests for internal secrecy do not suspend authorization controls.	Reduces the effectiveness of the confidentiality pretext, a recurring element of the modality.
Protect	Require out-of-band verification through a previously registered channel, initiated by the verifier, for transactions above a defined threshold.	Makes it harder for the adversary to control the confirmation channel.
Protect	Institute binding segregation of duties, with approval by two independent officers.	Raises the cost of the attack by requiring the simultaneous compromise of two decision-makers.
Protect	Adopt a code word or shared challenge, defined in advance and over a separate channel, for exceptional transactions.	Introduces a verification factor not derivable from public material.
Detect	Implement correlation rules for amount fragmentation, new destinations, and deviation from historical patterns.	May increase the probability of interruption during the execution window.
Respond	Activate the banking recovery channel as promptly as possible, per a target defined in the organization's policy, and institute a no-penalty policy for the timely reporting of suspicion or error.	May help reduce the risk of international dispersion of funds and the latency between suspicion and report.

Note. Compiled by the authors based on National Institute of Standards and Technology (2024), International Organization for Standardization (2022), and on the analysis of the case. The effects indicated are propositions derived from the analysis; the controls were neither implemented nor empirically evaluated in this study.

This article is published in Open Access under the Creative Commons Attribution 4.0 International (CC-BY) license, which allows unrestricted use, distribution and reproduction in any medium, provided the original author and source are credited.

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

None of the proposed controls depends on automated detection of synthetic content: the available detectors degrade outside their training domain and are targets of adversarial optimization, so that grounding the defense in detection would transfer the security of the process to a technological race unfavorable to the defender.

3.6 DISCUSSION OF THE RESULTS

The results converge with the literature as to the role of the principles of authority and scarcity, but add an element not sufficiently addressed: the mobilization of social proof through the simultaneous presence of multiple synthetic interlocutors — a qualitative, not quantitative, amplification. One falsified interlocutor requires the target to suspend a single verification; several converging interlocutors require suspending the hypothesis that such coordination is feasible, a hypothesis rendered obsolete by the near-zero marginal cost of additional synthesis. The findings also engage with Chesney and Citron (2019) on the degradation of the evidentiary authority of the audiovisual record: if video conferencing does not prove identity, current practices such as remote approvals, hiring interviews, and credential recovery through support desks operate on a compromised foundation.

From a risk management standpoint, the appropriate response is not necessarily technological: the asymmetry between generation and detection favors the adversary, whereas process controls remain effective regardless of the quality of the falsification, because they do not depend on distinguishing the true from the false but on requiring of the adversary control over channels they do not control. Finally, a revision of awareness programs is suggested: those oriented toward cue recognition tend toward obsolescence, while those oriented toward procedural discipline prove more robust, since they require the professional to act against their own perception.

4 - FINAL CONSIDERATIONS

This work examined, through a documentary holistic single-case study, the control gaps exploited by a social engineering attack assisted by generative artificial intelligence, taking as its object the January 2024 incident at the Hong Kong office of a British multinational engineering firm.

The attack chain was reconstructed in five phases from triangulated public evidence, showing that the decisive phase was not initial contact but the overcoming of the resistance legitimately expressed by the professional. The mapping according to MITRE ATT&CK revealed an atypically short chain, without any interaction with the organization's systems, which indicates the inadequacy of perimeter controls for this class of threat. Six control gaps were inferred from the available public evidence, in light of ISO/IEC 27001:2022, of which the most consequential concerns identity verification, and a set of compensating controls organized by the functions of the NIST Cybersecurity Framework 2.0 was proposed, all of them independent of automated detection of synthetic content.

This article is published in Open Access under the Creative Commons Attribution 4.0 International (CC-BY) license, which allows unrestricted use, distribution and reproduction in any medium, provided the original author and source are credited.

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

The theoretical contribution consists in the proposition that video conferencing should be reclassified as an unauthenticated channel, with the consequence that confirmation by video does not satisfy the out-of-band verification requirement. The practical contribution lies in the set of controls proposed, predominantly procedural and governance-related, whose adoption depends above all on a decision regarding the suppression of procedural exceptions; the study did not assess the costs, infrastructure, or operational impact of their implementation. The limitations already made explicit are acknowledged, in particular the absence of access to internal documentation, which gives the gaps the status of inference.

As a future agenda, four directions are suggested: a multiple-case study comparing incidents of synthetic impersonation across different sectors; a survey on the degree of adoption of out-of-band verification in Brazilian organizations; an experimental study of the comparative effectiveness of awareness programs oriented toward cues versus procedural discipline; and an analysis of the legal implications of synthetic impersonation under Brazilian law, as regards the liability of the agent who executes an undetectably falsified order and the treatment of the biometric data used to train the models in light of the General Personal Data Protection Law.

REFERENCES

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

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*.

Law No. 13,709, of August 14, 2018. (2018). *Lei Geral de Proteção de Dados Pessoais* [General Personal Data Protection Law]. Presidência da República. https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm

Magramo, K. (2024, May 17). 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

This article is published in Open Access under the Creative Commons Attribution 4.0 International (CC-BY) license, which allows unrestricted use, distribution and reproduction in any medium, provided the original author and source are credited.

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, May 17). 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.

This article is published in Open Access under the Creative Commons Attribution 4.0 International (CC-BY) license, which allows unrestricted use, distribution and reproduction in any medium, provided the original author and source are credited.