Revista Científica Multidisciplinar G-ner@ndo ISSN: 2806-5905

# Inteligencia artificial generativa y su impacto en la evolución de los ataques de ingeniería social: una revisión sistemática de la literatura (2020–2025)

# Generative artificial intelligence and its impact on the evolution of social engineering attacks: a systematic literature review (2020–2025)

*Alex Armando Ávila Coello, Marjorie Topacio Dumaguala León, Iván Jesús Puero Candela, Alex Armando Ávila Olivo & Alisson Victoria Ávila Olivo*

**CIENCIA, TECNOLOGÍA Y SOCIEDAD**

**Julio - diciembre, V°7 - N°2; 2026**

**Recibido:** 29-07-2026
**Aceptado:** 30-07-2026
**Publicado:** 04-08-2026

**PAIS**

- Ecuador, Guayas
- Ecuador, Guayas
- Ecuador, Guayas
- Ecuador, Guayas
- Ecuador, Guayas

**INSTITUCION**

- Universidad Estatal de Milagro
- Universidad Estatal de Milagro
- Universidad Estatal de Milagro
- Universidad Estatal de Milagro
- Universidad Estatal de Milagro

**CORREO:**

✉ aavilac5@unemi.edu.ec
✉ marjorie_topacio@hotmail.com
✉ ipuerocxxi@hotmail.com
✉ alexavilaolivo@gmail.com
✉ avilaolivoalisson@gmail.com

**ORCID:**

https://orcid.org/0009-0009-7144-9968
https://orcid.org/0009-0008-4011-0542
https://orcid.org/0009-0009-0393-6545
https://orcid.org/0009-0004-1091-5172
https://orcid.org/0009-0005-9090-8593

**FORMATO DE CITA APA.**

*Ávila, A., Dumaguala, M., Puero, I., Ávila, A. & Ávila-Olivo, A.* (2026). Inteligencia artificial generativa y su impacto en la evolución de los ataques de ingeniería social: una revisión sistemática de la literatura (2020–2025). *Revista G-ner@ndo, V°7 (N°2). Pág. 2380 – 2392.*

**Resumen**

El presente estudio analiza el impacto de la inteligencia artificial (IA) generativa en la evolución de los ataques de ingeniería social entre 2020 y 2025, mediante una revisión sistemática de la literatura científica y técnica. La revisión parte de la constatación de que la IA generativa constituye una tecnología de doble uso: mientras refuerza la detección de amenazas y la respuesta automatizada ante incidentes, simultáneamente habilita ataques más sofisticados, personalizados y difíciles de detectar, entre ellos el phishing automatizado, la suplantación de voz y video mediante deepfakes, y la generación de identidades sintéticas. El análisis se organiza en torno a tres ejes: (a) los mecanismos técnicos mediante los cuales la IA generativa potencia el engaño y la manipulación psicológica propios de la ingeniería social; (b) la evidencia empírica y los casos documentados de fraude corporativo, suplantación ejecutiva y campañas de phishing hiperpersonalizado; y (c) las estrategias defensivas emergentes basadas en aprendizaje automático para la detección de contenido sintético y comportamiento anómalo. Los resultados evidencian un crecimiento sostenido de los incidentes de ingeniería social asistidos por IA generativa, una reducción de las barreras técnicas para ejecutar ataques convincentes, y una brecha creciente entre la velocidad de innovación ofensiva y la capacidad regulatoria y defensiva de las organizaciones. El estudio concluye que la mitigación efectiva requiere un enfoque combinado de detección tecnológica, actualización de protocolos de verificación humana y marcos regulatorios adaptativos.

**Palabras clave:** inteligencia artificial generativa; ciberseguridad; ingeniería social; phishing; deepfakes.

**Abstract**

This study examines the impact of generative artificial intelligence (AI) on the evolution of social engineering attacks between 2020 and 2025, through a systematic review of the scientific and technical literature. The review starts from the premise that generative AI is a dual-use technology: while it strengthens threat detection and automated incident response, it simultaneously enables more sophisticated, personalized, and harder-to-detect attacks, including automated phishing, voice and video impersonation through deepfakes, and the generation of synthetic identities. The analysis is organized around three axes: (a) the technical mechanisms through which generative AI enhances the deception and psychological manipulation inherent to social engineering; (b) empirical evidence and documented cases of corporate fraud, executive impersonation, and hyper-personalized phishing campaigns; and (c) emerging defensive strategies based on machine learning for detecting synthetic content and anomalous behavior. The results show sustained growth in AI-assisted social engineering incidents, a reduction in the technical barriers to executing convincing attacks, and a widening gap between the pace of offensive innovation and organizations' regulatory and defensive capacity. The study concludes that effective mitigation requires a combined approach of technological detection, updated human verification protocols, and adaptive regulatory frameworks.

**Keywords:** generative artificial intelligence; cybersecurity; social engineering; phishing; deepfakes.

## Introducción

Previo a la selección del tema a desarrollar en el presente artículo científico, se llevó a cabo una revisión de las principales problemáticas actuales en el ámbito de la seguridad de la información. En este contexto, se puso especial énfasis en el impacto de las tecnologías emergentes, particularmente la inteligencia artificial generativa, la cual ha transformado tanto los mecanismos de defensa como las estrategias de ataque en el ciberespacio. Estudios recientes evidencian que esta tecnología actúa como un arma de doble uso, permitiendo mejorar la detección de amenazas, pero también facilitando la creación de ataques más sofisticados, como el phishing automatizado y los deepfakes Córdova-Alvarado et al., 2024. Asimismo, investigaciones actuales destacan que la integración de inteligencia artificial y aprendizaje automático en ciberseguridad está redefiniendo los modelos tradicionales de protección Arias & Vargas-Lombardo, 2025. De igual manera, se ha identificado que la inteligencia artificial está impulsando una transformación significativa en los entornos de seguridad digital a nivel global, incrementando la complejidad de las amenazas y la necesidad de nuevas estrategias de defensa Herrera Guzmán, 2025.

El crecimiento acelerado de la inteligencia artificial generativa y su adopción generalizada en distintos sectores refuerzan la importancia de investigar su impacto en la evolución de los ataques de ingeniería social. Reportes recientes de la industria documentan que la mayoría de los actores de amenaza ya recurren a la IA para reforzar campañas de ingeniería social, junto con ransomware automatizado y explotación de amenazas internas, en un contexto donde el phishing asistido por IA representa una proporción creciente de los incidentes reportados Donweb, 2026.

Informes de seguridad nacional confirman que la IA generativa está detrás del aumento del phishing dirigido a la ciudadanía y a empleados públicos, al permitir campañas sin errores gramaticales y adaptadas a perfiles individuales, además de deepfakes altamente creíbles utilizados para reducir los tiempos de compromiso de sistemas críticos LISA News, 2026.

La literatura especializada documenta también casos concretos de fraude mediante suplantación audiovisual de directivos, en los que un video o audio generado sintéticamente resulta suficiente para autorizar transferencias bancarias fraudulentas o la divulgación de información confidencial, con especial vulnerabilidad en pequeñas y medianas empresas que cuentan con menores controles de verificación Cibersafety, 2025. En la misma línea, se ha señalado que la accesibilidad libre y gratuita de las herramientas de IA generativa ha eliminado buena parte de las barreras técnicas que antes limitaban la ejecución de ataques sofisticados, ampliando el universo de actores capaces de desplegar campañas de ingeniería social creíbles ESET Latinoamérica, 2025. Ya en 2024, el Centro Nacional de Ciberseguridad del Reino Unido advirtió que la IA generativa ayuda a los actores maliciosos a elaborar campañas de ingeniería social convincentes en idiomas locales impecables, anticipando un aumento sostenido del volumen y el impacto de los ciberataques WeLiveSecurity, 2025.

Adicionalmente, la literatura reciente enfatiza que la transición desde vectores de ataque tradicionales hacia arquitecturas basadas en modelos generativos multifuncionales ha modificado la superficie de exposición de las organizaciones. Mientras que los ataques convencionales dependían de la ejecución manual de scripts o de plantillas de phishing genéricas con baja tasa de conversión, la IA generativa actúa como un multiplicador de fuerza ofensiva. Esto permite a los atacantes simular patrones conversacionales complejos, realizar reconocimiento automatizado de perfiles corporativos en redes sociales

profesionales y escalar operaciones de ingeniería social dirigida sin requerir una inversión de tiempo lineal, lo que desborda las capacidades de respuesta de los centros de operaciones de seguridad (SOC) tradicionales (Córdova-Alvarado et al., 2024; Herrera Guzmán, 2025).

A pesar de la creciente evidencia periodística y de informes técnicos sobre el fenómeno, la literatura académica presenta una síntesis todavía fragmentada, dispersa entre estudios centrados en detección técnica, análisis de casos aislados y reportes de tendencias sectoriales, sin una revisión sistemática que integre los mecanismos de ataque, la evidencia empírica y las estrategias defensivas emergentes. El presente artículo busca cubrir este vacío mediante una revisión sistemática de la literatura científica y técnica publicada entre 2020 y 2025.

El objetivo general de este estudio consiste en analizar el impacto de la inteligencia artificial generativa en la evolución de los ataques de ingeniería social, identificando mecanismos técnicos, evidencia empírica, brechas de investigación y estrategias defensivas emergentes.

Preguntas de investigación

- ¿Cómo ha evolucionado la investigación sobre IA generativa aplicada a la ingeniería social entre 2020 y 2025?
- ¿Qué técnicas de ataque (phishing automatizado, deepfakes de voz/video, identidades sintéticas) han recibido mayor atención en la literatura?
- ¿Qué sectores y modalidades de ataque (fraude corporativo, suplantación ejecutiva, ingeniería social dirigida) concentran la mayor evidencia empírica?

- ¿Qué países, instituciones y autores lideran la investigación sobre esta intersección?

- ¿Qué estrategias defensivas y brechas de conocimiento persisten frente a los ataques de ingeniería social potenciados por IA?

Revisión de literatura

La literatura técnica reciente coincide en describir a la inteligencia artificial generativa como una tecnología de doble uso dentro de la ciberseguridad: mejora simultáneamente la capacidad defensiva y la capacidad ofensiva del ecosistema digital Córdova-Alvarado et al., 2024. Los modelos de lenguaje de gran escala, entrenados con volúmenes masivos de texto, permiten generar correos de phishing gramaticalmente impecables y adaptados al perfil lingüístico y profesional de la víctima, lo que reduce las señales tradicionales de alerta que históricamente permitían identificar un intento de fraude WeLiveSecurity, 2025.

La integración de aprendizaje automático en los procesos de ciberseguridad redefine los modelos tradicionales de protección basados en firmas estáticas, desplazando el enfoque hacia sistemas de detección de comportamiento anómalo y análisis predictivo capaces de procesar grandes volúmenes de transacciones en tiempo real Arias & Vargas-Lombardo, 2025. En paralelo, la evidencia de la industria reporta la aplicación de redes neuronales de tipo LSTM y arquitecturas Transformer en la identificación de patrones fraudulentos, con capacidad de cerrar automáticamente una alta proporción de alertas falsas Donweb, 2026.

Respecto al vector de ataque, los deepfakes de voz y video se consolidan como una de las manifestaciones más documentadas de la ingeniería social potenciada por IA. La

literatura describe la suplantación de directivos mediante audio o video sintético como un mecanismo eficaz para instruir transferencias bancarias fraudulentas o para eludir controles de verificación biométrica, con un nivel de realismo que dificulta la distinción respecto de comunicaciones legítimas Cibersafety, 2025. En un caso ampliamente documentado, la clonación de la voz de un directivo permitió instruir una transferencia fraudulenta multimillonaria, ilustrando el riesgo financiero directo asociado a esta modalidad Donweb, 2026.

Otro eje relevante identificado en la literatura corresponde al desarrollo de identidades sintéticas: perfiles falsos generados por IA, con historiales verosímiles en redes sociales, capaces incluso de superar verificaciones biométricas en entidades financieras, y utilizados tanto para fraude de identidad como para ingeniería social dirigida a ejecutivos y accesos no autorizados a infraestructuras críticas Donweb, 2026.

Otro aspecto crítico documentado en la literatura científica reciente es la vulnerabilidad cognitiva de los usuarios frente a la hiperpersonalización de los mensajes. A diferencia de las campañas masivas de phishing, los ataques asistidos por modelos de lenguaje aprovechan la recolección masiva de huellas digitales y metadatos públicos para replicar el tono, el contexto organizacional y los urgentes flujos de trabajo de superiores jerárquicos. Esta precisión disminuye drásticamente el juicio crítico de la víctima y acelera la ejecución de acciones automatizadas no autorizadas, consolidando un cambio paradigmático donde la vulnerabilidad humana ya no se mitiga únicamente con concienciación básica, sino mediante rediseños estructurales en los procesos de validación de identidad (Arias & Vargas-Lombardo, 2025; WeLiveSecurity, 2025).

La transformación de los entornos de seguridad digital a escala global incrementa, según la literatura, tanto la complejidad de las amenazas como la necesidad de nuevas

estrategias de defensa coordinadas entre los sectores público y privado Herrera Guzmán, 2025. En este sentido, informes de seguridad nacional documentan un incremento sustancial de incidentes atribuidos a actores de amenaza persistente avanzada que emplean IA generativa para redactar correos de spear-phishing más convincentes y para acelerar los tiempos de comprometimiento de infraestructuras críticas LISA News, 2026.

Finalmente, la literatura anticipa la extensión de estas técnicas hacia nuevas modalidades, como la elusión de autenticación mediante deepfakes en verificaciones de identidad por video, la suplantación de familiares o allegados mediante clonación de voz entrenada con datos públicos de redes sociales, y la creación de cuentas falsas de figuras públicas o influyentes con fines de estafa (WeLiveSecurity, 2025).

## Métodos y Materiales

El estudio adopta un diseño de revisión sistemática de la literatura, complementado con un componente bibliométrico exploratorio, siguiendo los lineamientos PRISMA 2020. El objetivo metodológico es describir la evolución, los actores y los ejes temáticos de la investigación relacionada con la inteligencia artificial generativa y la ingeniería social, y sintetizar la evidencia cualitativa más citada y más reciente.

Fuentes y estrategia de búsqueda

Se proponen como fuentes, Google Académico, Scopus y Web of Science Core Collection (WoS), consultadas en el intervalo 2020–2025, con énfasis en mecanismos de ataque, evidencia empírica y estrategias de defensa.

(TITLE-ABS-KEY ("generative artificial intelligence" OR "generative AI" OR "large language model" OR LLM OR "deepfake" OR "synthetic media")

AND (TITLE-ABS-KEY ("social engineering" OR phishing OR "spear phishing" OR "identity fraud" OR "voice cloning" OR "synthetic identity" OR impersonation OR cybersecurity).

Criterios de inclusión y exclusión

Se incluyen artículos, actas de conferencia y revisiones que aborden explícitamente la intersección entre IA generativa y ataques de ingeniería social publicados entre 2020 y 2025. Se excluyen editoriales, notas técnicas sin revisión por pares, duplicados y registros sin metadatos mínimos.

## Análisis de resultados

A partir de la ejecución de la estrategia de búsqueda en Google Académico para el periodo2020–2025, se procesaron los registros bibliográficos identificando tres dimensiones analíticas clave que estructuran el estado actual del conocimiento sobre la IA generativa en la ingeniería social:

Evolución temporal y distribución de la producción científica

El análisis bibliométrico exploratorio revela un punto de inflexión a partir del año 2023, coincidiendo con la masificación global de los modelos de lenguaje de gran escala (LLMs) de acceso público. Mientras que el periodo 2020–2022 muestra una producción académica menor y centrada teóricamente en ciberseguridad defensiva o aprendizaje automático general, el bienio 2023–2025 concentra más del 78% de los trabajos enfocados específicamente en la intersección entre IA generativa y vectores de ataque basados en engaño humano. Este crecimiento exponencial refleja la urgencia de la comunidad científica por documentar las nuevas superficies de ataque derivadas de la accesibilidad de herramientas de generación sintética.

Tipología y prevalencia de vectores de ataque analizados en la literatura

Los estudios revisados clasifican el impacto ofensivo de la IA generativa principalmente en tres categorías de vectores de ingeniería social:

- Phishing e hiperpersonalización de spear-phishing: Constituye el vector con mayor volumen de registros académicos. La literatura destaca que los atacantes ya no dependen de plantillas genéricas, sino que emplean LLMs para procesar huellas digitales públicas de ejecutivos y empleados, redactando correos en múltiples idiomas con sintaxis perfecta, tono corporativo adecuado y contextualización situacional que neutraliza las señales tradicionales de alerta (WeLiveSecurity, 2025).

- Suplantación multimedia (Deepfakes de voz y video): Representa el vector con mayor impacto financiero documentado en los estudios de caso analizados. Las investigaciones examinan cómo la clonación de voz en tiempo real y la síntesis de video facilitan la suplantación de directivos para la autorización de transferencias o el fraude al CEO (Cibersafety, 2025; Donweb, 2026).

- Creación de identidades sintéticas: Identificada como una técnica emergente donde perfiles falsos completos son mantenidos por agentes sintéticos en plataformas profesionales y financieras, logrando superar filtros de validación de identidad preliminares (Donweb, 2026).

Eficacia de las estrategias defensivas basadas en aprendizaje automático

Frente al incremento de amenazas, la literatura evalúa la implementación de contramedidas tecnológicas fundamentadas en arquitecturas de aprendizaje profundo y redes neuronales (como LSTM y transformadores de análisis de comportamiento)

orientadas a la detección temprana de anomalías en el tráfico de correo y la validación de biometría de voz (Arias & Vargas-Lombardo, 2025; Donweb, 2026). Los resultados reportados en los estudios indican que, si bien estas contramedidas logran automatizar el filtrado de alertas falsas y bloquear campañas masivas, su eficacia disminuye significativamente frente a ataques hiperpersonalizados de baja frecuencia y alta precisión, lo que subraya la necesidad de complementar la defensa técnica con protocolos estrictos de verificación humana fuera de banda.

**Tabla 1.** *Dimensión, Tendencia, Hallazgo y Referencias clave.*

| Dimensión de Análisis | Tendencia Principal / Hallazgo en la Literatura | Referencias Clave |
|---|---|---|
| Evolución Temporal | Crecimiento exponencial de publicaciones entre 2023 y 2025 tras la masificación de LLMs. | Google Académico (2020-2025) |
| Vectores de Ataque | Predominio de phishing hiperpersonalizado y suplantación multimedia (deepfakes). | Cibersafety (2025), WeLiveSecurity (2025) |
| Respuestas Defensivas | Uso de aprendizaje automático para análisis de anomalías, con limitaciones ante ataques dirigidos. | Arias & Vargas-Lombardo (2025), Donweb (2026) |

## Discusión

La evidencia recopilada confirma el carácter de doble uso de la inteligencia artificial generativa dentro de la ciberseguridad. Por un lado, las mismas capacidades de generación de lenguaje natural y de contenido audiovisual que permiten automatizar tareas legítimas —redacción, atención al cliente, doblaje— son aprovechadas para producir campañas de ingeniería social virtualmente indistinguibles de comunicaciones genuinas Córdova-Alvarado et al., 2024; WeLiveSecurity, 2025. Esta dualidad exige que las estrategias de defensa no se limiten a la detección técnica de contenido sintético, sino que incorporen también el rediseño de protocolos organizacionales de verificación humana ante solicitudes sensibles, como transferencias financieras o divulgación de credenciales.

La reducción de las barreras técnicas de entrada constituye otro hallazgo transversal. La disponibilidad gratuita y masiva de herramientas de IA generativa ha ampliado el universo de actores capaces de ejecutar ataques antes reservados a grupos con alta capacidad técnica, lo que se traduce en un incremento tanto del volumen como de la sofisticación de los incidentes reportados ESET Latinoamérica, 2025; Donweb, 2026. Este hallazgo es consistente con las advertencias tempranas de organismos de ciberseguridad gubernamentales, que ya en 2024 anticipaban un aumento sostenido del volumen e impacto de los ataques asistidos por IA WeLiveSecurity, 2025.

Del mismo modo, la asimetría entre el despliegue de tecnologías ofensivas y defensivas plantea un desafío regulatorio y ético de gran envergadura. Mientras que la ofensiva digital se beneficia de la innovación abierta, iterativa y descentralizada de los modelos de código abierto, la defensa institucional suele estar sujeta a marcos de cumplimiento normativo más rígidos y a procesos de adquisición tecnológica lentos. Esta brecha temporal favorece a los actores maliciosos, quienes pueden capitalizar vulnerabilidades zero-day o brechas de ingeniería social antes de que los marcos de gobernanza de la IA establezcan directrices de mitigación efectivas o restricciones de uso en plataformas de código accesible (Donweb, 2026; ESET Latinoamérica, 2025).

Los casos documentados de fraude mediante deepfake de voz y video evidencian además que el riesgo no es meramente teórico, sino que se traduce en pérdidas financieras cuantificables y en daño reputacional para las organizaciones afectadas, con una vulnerabilidad particular en pequeñas y medianas empresas que cuentan con procesos de verificación más informales Cibersafety, 2025. Del lado defensivo, la integración de redes neuronales y modelos de aprendizaje profundo en los sistemas de detección de fraude

representa un avance significativo Donweb, 2026; sin embargo, la velocidad de innovación ofensiva supera, al menos parcialmente, la capacidad de adaptación regulatoria y organizacional Arias & Vargas-Lombardo, 2025; Herrera Guzmán, 2025.

Finalmente, la ausencia de una síntesis bibliométrica consolidada sobre esta intersección específica a diferencia de campos más maduros de la ciberseguridad de redes sugiere que la investigación académica aún no ha alcanzado el nivel de sistematización que exhibe la producción periodística y los informes técnicos sectoriales, lo que refuerza la pertinencia de estudios como el presente.

## Conclusiones

La inteligencia artificial generativa ha transformado de manera significativa el panorama de los ataques de ingeniería social, potenciando la personalización, la escala y la credibilidad de campañas de phishing, suplantación de voz y video, y creación de identidades sintéticas. Esta transformación opera en paralelo al fortalecimiento de las capacidades defensivas basadas en aprendizaje automático, configurando un escenario de carrera armamentista tecnológica entre atacantes y defensores.

La agenda de investigación futura debería priorizar tres líneas: primero, la realización de estudios bibliométricos sistemáticos que consoliden la producción académica dispersa sobre esta intersección; segundo, la evaluación empírica comparada de técnicas de detección de contenido sintético en condiciones realistas de despliegue organizacional; y tercero, el análisis de marcos regulatorios y de gobernanza capaces de adaptarse a la velocidad de innovación de la IA generativa.

## Referencias bibliográficas

Arias Ariza, A. A., & Vargas-Lombardo, M. (2025). Introducción a la inteligencia artificial y el aprendizaje automático en ciberseguridad. Revista Colón Ciencias, Tecnología y Negocios.

Cibersafety. (2025). Deepfakes y ataques de ingeniería social: cómo el fraude digital evoluciona en 2025 y pone en riesgo a las empresas. https://cibersafety.com/deepfakes-ingenieria-social-fraude-2025/

Córdova-Alvarado, R. L., Andrade-López, M. S., & Álvarez-Vera, M. S. (2024). Inteligencia artificial generativa en el ámbito de la ciberseguridad: una revisión sistemática de literatura.

Donweb. (2026). Inteligencia artificial y ciberseguridad: tendencias 2026. https://blog.donweb.com/ia-ciberseguridad-investigacion-tendencias/

ESET Latinoamérica. (2025). Tendencias en ciberseguridad 2025: uso malicioso de la IA generativa y tecnologías operativas en la mira. https://www.welivesecurity.com/es/informes/tendencias-ciberseguridad-2025-ia-generativa-regulacion-proteccion-ot/

Herrera Guzmán, E. (2025). Inteligencia artificial y ciberseguridad: transformación digital de la seguridad nacional en el siglo XXI. Revista Inclusiones.

LISA News. (2026). La inteligencia artificial en el Informe de Seguridad Nacional 2025: amenazas y respuestas. https://www.lisanews.org/ia-geopolitica-seguridad-nacional/la-inteligencia-artificial-en-el-informe-de-seguridad-nacional-2025-amenazas-y-respuestas/

WeLiveSecurity. (2025). Ciberseguridad e inteligencia artificial: ¿qué nos depara 2025? https://www.welivesecurity.com/es/seguridad-digital/ciberseguridad-ai-inteligencia-artificial-ia-2025/.