ARTICULO ORIGINAL

Nuevas modalidades de fraude digital, phishing, deepfakes e ingeniería social y su relación con el delito de estafa

Recibido: 10/02/2026
Aceptado: 27/03/2026
Publicado: 01/04/2026

Andrés Alejandro Zuleta Araque
https://orcid.org/0009-0008-2925-2267
Universidad Estatal Península de Santa Elena, Ecuador
azuleta@upse.edu.ec
Magíster en Derecho Penal; Especialista Superior y Máster en Derecho Procesal. Abogado. Docente de la carrera de derecho de la Universidad Estatal Península de Santa Elena

Wilfrido Wasbrum Tinoco
https://orcid.org/0000-0002-6172-5633
Universidad Estatal Península de Santa Elena, Ecuador
wwasbrum@upse.edu.ec
Magister en Derecho Penal, Docente de Derecho penal, Universidad Península de Santa Elena

Resumen

La estafa es un delito que tiene una multiplicidad de modalidades de cometimiento, con el auge de las tecnologías de la información este delito se ha masificado en lo digital, especialmente las modalidades donde el delincuente busca captar información del usuario para el fraude que en los últimos 4 años se han visto ampliamente potenciadas por la implementación de la inteligencia artificial. Este artículo tiene como objetivo analizar las nuevas modalidades de estafa que surgen a partir del desarrollo de inteligencia artificial, este estudio se centra especialmente phishing automatizado, los deepfakes, la clonación de voz, la ingeniería social, las estafas románticas y los fraudes financieros digitales. El estudio se desarrolló bajo un enfoque cualitativo, de tipo documental-descriptivo, mediante la revisión de literatura académica publicada entre 2020 y 2026. Los resultados de este estudio evidencian que la inteligencia artificial generativa ha incrementado la capacidad de los ciberdelincuentes para personalizar, automatizar y masificar el engaño, generando comunicaciones fraudulentas más creíbles y difíciles de detectar. Sin embargo, bajo un análisis jurídico de este tipo de conductas, estas nuevas modalidades de estafa no suponen un cambio en la estructura básica del delito de estafa, sino que modifican los medios que se emplean para cometer esta. En consecuencia, el engaño, el error, la disposición patrimonial, el perjuicio y el ánimo de lucro continúan siendo elementos esenciales para su configuración, aunque en estos casos estos se dan en espacios digitales. Se concluye que el principal reto para el derecho penal contemporáneo consiste en adaptar sus criterios de interpretación, investigación y prueba a escenarios donde la inteligencia artificial permite simular identidades, relaciones de confianza y comunicaciones aparentemente legítimas.

Palabras clave: estafa digital; inteligencia artificial; phishing; deepfakes; ingeniería social; derecho penal.

Abstract

Fraud is a crime with multiple forms of commission. With the rise of information technologies, this crime has become widespread in the digital realm, especially in methods where the perpetrator seeks to capture user information for fraudulent purposes. These methods have been significantly enhanced in the last four years by the implementation of artificial intelligence. This article aims to analyze the new forms of fraud that have emerged from the development of artificial intelligence. This study focuses particularly on automated phishing, deepfakes, voice cloning, social engineering, romance scams, and digital financial fraud. The study was conducted using a qualitative, documentary-descriptive approach, through a review of academic literature published between 2020 and 2026. The results of this study demonstrate that generative artificial intelligence has increased cybercriminals' ability to personalize, automate, and scale up deception, generating more credible and difficult-to-detect fraudulent communications. However, from a legal perspective, these new forms of fraud do not represent a change in the basic structure of the crime of fraud, but rather modify the means employed to commit it. Consequently, deception, error, transfer of assets, harm, and the intent to profit remain essential elements for its definition, although in these cases they occur in digital spaces. It is concluded that the main challenge for contemporary criminal law lies in adapting its criteria for interpretation, investigation, and evidence to scenarios where artificial intelligence allows for the simulation of identities, relationships of trust, and seemingly legitimate communications.

Key words: digital fraud; artificial intelligence; phishing; deepfakes; social engineering; criminal law.

Introducción

Desde 2020, el fraude digital ha evolucionado hacia formas de engaño asistidas por inteligencia artificial (IA), en varias modalidades especialmente mediante phishing, spear phishing personalizado, deepfakes, clonación de voz, suplantación de identidad, ingeniería social multicanal y generación masiva de mensajes fraudulentos. Estas modalidades conservan lo esencial del tipo penal de la estafa (engaño, error, disposición patrimonial y perjuicio), pero varia de forma radical en los medios que utiliza para cometer el delito y la capacidad de masificar el delito debido a emplear la inteligencia artificial como herramienta preferentemente utilizada para la comisión del delito en el campo de lo digital. Esto debido a que a medida que la tecnología evoluciona, sus capacidades aumentan y, sobre todo, sus costes se reducen drásticamente, como demuestran en su trabajo Schmitt y Flechais (2024) “Inteligencia artificial generativa en ingeniería social y phishing”. Estas previsiones sitúan la combinación de ingeniería social e IA en una posición emergente, ya que los ataques de ingeniería social suelen ofrecer una forma menos compleja de acceder a un sistema, sin necesidad de sortear contramedidas tecnológicas complejas.

Actualmente, el delito de estafa en el mundo digital se ve potenciada por la implementación de la inteligencia artificial generativa, lo que es una amenaza para la sociedad actual que se encuentra ampliamente interconectada por la implementación de internet en muchos aspectos de la vida cotidiana. El avance de los sistemas de IA ocasiona que estos cada día desarrollen una capacidad de imitar la capacidad de comunicación humana y a través de esto ocasionar que el receptor perciba confianza de un supuesto emisor que sería humano, por lo que esta situación de una IA desarrollada con estas nuevas capacidades abre nuevos retos para la ingeniería social y phishing mediante el uso de estas herramientas. (Schmitt and Flechais, 2024) El desarrollo de esta tecnología abre nuevas posibilidades para el cometimiento de delitos en el ciberespacio, si bien estos existen desde la masificación del internet en el mundo, la llegada de la inteligencia artificial ocasiona que esto delitos tenga una masificación que está relacionada directamente a la aparición de la IA.

Los modelos de lenguaje que conforman lo que conocemos como inteligencia artificial sirven como una herramienta para que usuarios maliciosos puedan diseñar correos electrónicos de tipo phishing a partir de un poco de información de un particular. Estos nuevos tipos de correos contrastan con los correos electrónicos de phishing tradicionales que los hackers diseñan manualmente utilizando reglas generales extraídas de la experiencia, por lo que se evidencia que el avance de la tecnología en IA facilita esta modalidad de phishing debido a que mediante la aplicación de la inteligencia artificial se puede generar de forma masiva mensajes que luego serán distribuidos bajo la modalidad de phishing, siendo este último el más común de los ciberdelitos. (Heiding et al., 2023)

El rápido crecimiento de las redes sociales y las plataformas de mensajería ha aumentado drásticamente la exposición de los usuarios a las estafas en línea. Estos ataques, que van desde correos electrónicos de phishing y llamadas de suplantación de identidad hasta mensajes directos fraudulentos, explotan la información personal disponible públicamente y utilizan técnicas de manipulación psicológica como la urgencia, el miedo y las señales de autoridad para engañar a las personas y lograr que revelen datos confidenciales, si bien muchos individuos consideran que no caerían en este tipo de estafa debido a que esto es ampliamente conocido en la actualidad por los usuarios más jóvenes, está claro que el grupo más vulnerable a este tipo de afectaciones son los usuarios que no son nativos digitales o tienen barreras de acceso a este tipo de herramientas digitales (Hossain et al.,., 2025)

Los informes sobre fraudes y ciberdelitos facilitados por IA están en aumento en los últimos años. Según se informa, los atacantes utilizan la IA para obtener empleos remotos con identidades falsas, crear perfiles de las víctimas y diseñar sofisticadas campañas de phishing, cuestión que incluso es reconocida actualmente por sus propios desarrolladores como un riesgo actual inherente al uso de estas herramientas (Anthropic, 2025a; OpenAI, 2025a). Según un informe de Lookout Security, “2022 marcó la tasa más alta de encuentros de phishing móvil”. Desde entonces, la situación solo ha empeorado. InfoSecurity Magazine informa que “los encuentros de phishing móvil han aumentado cada trimestre desde el segundo trimestre de 2020”, con un asombroso aumento del 202% en los mensajes de phishing en general y un dramático aumento del 703% en los ataques de phishing de credenciales en la segunda mitad de 2024.(Denisenko et al., 2026) El impacto en las víctimas es potencialmente significativo: datos recientes del Reino Unido muestran que el fraude romántico cuesta 106 millones de libras esterlinas al año, el fraude de identidad representa el 59 % de todos los casos denunciados y el fraude del CEO causa pérdidas promedio superiores a las 10 000 libras esterlinas. (Mai et al., 2026)

A pesar del enfoque en mejorar la concienciación sobre ciberseguridad por parte de todos los gobiernos y organizaciones, el número de ciberataques ha aumentado significativamente en los últimos años, lo que ha provocado enormes pérdidas financieras y cuyos riesgos se extienden por todo el mundo. Esto se debe a las técnicas empleadas en los ciberataques, que principalmente buscan explotar la vulnerabilidad psicológica de los usuarios, mismo que es el eslabón más débil de cualquier sistema de defensa contra fraudes electrónicos. (Jabir et al., 2025) Si bien existen actualmente ciertos tipos de herramientas que sirven para poder detectar ciberataques, estas cuando funcionan de manera automática tienen algún precio en el mercado por lo cual no están al alcance de todos los usuarios, lo cual determina que en la mayoría de los casos el usuario solamente depende de su criterio para poder decidir si un mensaje que recibe es de origen malicioso o no.

El presente estudio se desarrolló desde un enfoque cualitativo, de tipo documental y descriptivo. En términos sencillos, lo que se hizo fue revisar bibliografía reciente relacionada con las nuevas formas de estafa digital que han surgido o se han perfeccionado con el uso de la inteligencia artificial. Para ello, se tomó como referencia el período comprendido entre 2020 y 2026, ya que el objetivo era trabajar con literatura actual y vinculada a fenómenos recientes como el phishing generado por IA, los deepfakes, la ingeniería social automatizada, los fraudes financieros digitales, las estafas románticas y el uso de modelos de lenguaje de gran escala en escenarios de engaño y ciberdelincuencia.

El método empleado fue documental-bibliográfico, complementado con un análisis hermenéutico-jurídico, ya que en este estudio se revisaron publicaciones especializadas con el propósito de relacionar sus aportes con los elementos del tipo penal de la estafa en la legislación ecuatoriana. El diseño de la investigación fue no experimental, de carácter transversal-retrospectivo, puesto que en este trabajo no se manipularon variables, por lo que principalmente este trabajo se basa en análisis de documentos publicados entre 2020 y 2026. El tipo de investigación aplicado en este trabajo fue documental y el nivel empleado fue descriptivo-analítico, debido a que se identificaron, describieron y analizaron las principales modalidades de estafa asistida por inteligencia artificial, tales como phishing automatizado, deepfakes, ingeniería social, fraudes financieros digitales, romance scams y uso de modelos de lenguaje de gran escala.

La población estuvo conformada por publicaciones académicas, artículos científicos, revisiones y estudios especializados relacionados con inteligencia artificial, fraude digital, estafa, phishing, deepfakes, ingeniería social y ciberdelincuencia. Por otro lado, se excluyeron los documentos que no tenían autoría verificable, las publicaciones meramente periodísticas, los textos duplicados, las fuentes que no guardaban una relación directa con la estafa o el fraude digital, y los trabajos anteriores al año 2020. Con base en este proceso, la muestra final estuvo integrada por 20 documentos académicos seleccionados de manera intencional, considerando su pertinencia temática, actualidad, autoría verificable y relación directa con el objeto de estudio.

A los documentos seleccionados se les aplico una técnica de análisis documental por medio de la implementación de una matriz documental, en ella se registraron datos como autor, año, título, fuente, modalidad de fraude estudiada, tecnología utilizada, principales hallazgos y posible relación con los elementos jurídicos de la estafa. Esta forma de organización permitió ordenar la literatura en varios ejes temáticos: phishing generado por inteligencia artificial, deepfakes y suplantación de identidad, fraudes financieros automatizados, estafas románticas asistidas por modelos de lenguaje y uso de agentes conversacionales en contextos de ciberdelincuencia.

Para ubicar inicialmente la literatura se utilizó una estrategia de búsqueda apoyada en herramientas de inteligencia artificial generativa. Sin embargo, es importante aclarar que estas herramientas no fueron usadas como fuente científica ni doctrinal, sino únicamente como un apoyo exploratorio para identificar posibles autores, títulos, palabras clave y líneas temáticas relevantes. Después de esa primera búsqueda, cada referencia fue revisada y verificada en fuentes académicas y editoriales especializadas, como revistas científicas, repositorios de preprints, bases de datos académicas, páginas editoriales y registros DOI. Es decir, la inteligencia artificial sirvió como una especie de punto de partida para orientar la búsqueda, pero no como fundamento académico del estudio.

Los criterios de inclusión permitieron seleccionar trabajos publicados desde el año 2020 en adelante, especialmente artículos científicos, capítulos, preprints académicos e investigaciones especializadas relacionadas con fraude digital, estafa, phishing, deepfakes, ingeniería social, modelos de lenguaje, agentes de inteligencia artificial y ciberdelincuencia. También se consideró necesario que los textos tuvieran autoría identificable y pudieran ser verificados mediante DOI, revista científica, repositorio académico o plataforma editorial. Además, se priorizaron aquellos estudios que permitían conectar sus hallazgos con las nuevas formas de engaño patrimonial que hoy resultan relevantes para el derecho penal.

Finalmente, la información fue interpretada desde una perspectiva jurídico-penal. Es decir, los aportes tecnológicos y criminológicos encontrados en la literatura fueron relacionados con los elementos tradicionales de la estafa: el engaño, el error, la disposición patrimonial, el perjuicio y el ánimo de lucro. A partir de esto, se pudo evidenciar que las nuevas modalidades digitales no eliminan la estructura clásica del delito de estafa, sino que más bien modifican los medios mediante los cuales se produce el engaño. En otras palabras, cambia la forma en que se comete el delito, pero su núcleo esencial sigue siendo el mismo.

El papel de la IA en los fraudes digitales

La inteligencia artificial generativa ha incrementado las capacidades de los ciberdelincuentes para automatizar, personalizar y masificar comunicaciones fraudulentas, especialmente en modalidades como el phishing, la ingeniería social y el fraude financiero digital (Denisenko et al., 2026; Heiding et al., 2023; Jabir et al., 2025; Schmitt & Flechais, 2024).Los recientes avances en inteligencia artificial en todo el mundo, particularmente en el ámbito de los modelos de lenguaje a gran escala (LLM), han dado como resultado sistemas potentes y versátiles de doble uso. Esta inteligencia puede emplearse para una amplia variedad de tareas beneficiosas, pero también puede utilizarse para actividades delictivas en línea. Los LLM puede ser utilizados para el spear phishing, que es una forma de ciberdelincuencia que consiste en manipular a las víctimas para que revelen información confidencial. (Hazell 2023) El phishing asistido por inteligencia artificial representa una evolución

del phishing tradicional, debido a que los modelos de lenguaje permiten generar mensajes más creíbles, personalizados y adaptados al perfil de la víctima, lo que dificulta su detección por parte de los usuarios (Denisenko et al., 2026; Hazell, 2023; Heiding et al., 2023; Jabir et al., 2025).

Autores como Badhe (2025) aseveran que los modelos de lenguaje a gran escala han demostrado una fluidez y capacidad de razonamiento impresionantes, pero su potencial de mal uso ha generado una creciente preocupación. La reciente disponibilidad de una potente inteligencia artificial como herramienta cotidiana ha impulsado una nueva ola de técnicas de ataque, especialmente en el ámbito de la Ingeniería Social. La posibilidad de generar multitud de plantillas diferentes en cuestión de segundos para llevar a cabo un ataque de ingeniería social reduce la barrera de entrada para los posibles ciberdelincuentes. (Matekas et al., 2025)

Esto debido a que anteriormente para poder generar un deepfake la barrera de acceso era limitada debiendo el delincuente manejar conocimientos de edición de imagen y videos de nivel semi profesional, cuestión que actualmente con la IA no existe debido a su versatilidad para generar información con simples comandos, por lo que la incorporación de inteligencia artificial en escenarios de fraude digital reduce las barreras técnicas para los delincuentes, ya que permite generar mensajes, audios, imágenes o interacciones falsas sin requerir conocimientos avanzados de programación, edición audiovisual o manipulación informática (Badhe, 2025; Denisenko et al., 2026; Matecas et al., 2025; Schmitt & Flechais, 2024).

Mientras Schmitt y Flechais (2024) enfatizan que la inteligencia artificial generativa potencia la ingeniería social y el phishing, Chlasta (2025) advierte que esta misma tecnología tiene una dimensión dual, ya que puede ser usada tanto para atacar como para defender sistemas digitales, siendo un factor significativo que configura el ecosistema de ciberseguridad actual dado que, por un lado, sirve como una herramienta eficaz para apoyar la protección de los sistemas de información y, por otro, puede ser explotada como vector de ataque por actores maliciosos. Por su parte, Jiang (2024) refuerza esta segunda perspectiva al analizar el uso de modelos de lenguaje para la detección de estafas.

Siendo este último el uso que se le da por parte de ciberdelincuentes que explotan la capacidad de la inteligencia artificial para generar diferentes tipos de fraudes, por lo que el lanzamiento de la inteligencia artificial fue un acontecimiento que ha marcado de forma sostenida el incremento de fraudes digitales que puede ser implementados y facilitados por medio la aplicación de este tipo de herramientas. La literatura reciente coincide en que la inteligencia artificial posee un carácter dual, pues puede ser utilizada tanto por ciberdelincuentes para perfeccionar ataques de phishing, deepfakes o fraudes financieros, como por instituciones y usuarios para detectar patrones sospechosos y prevenir estafas digitales (Chlasta, 2025; Hossain et al., 2025; Jiang, 2024; Schmitt & Flechais, 2024).

Jiang, (2024) sostiene que las estafas son prácticas engañosas diseñadas para explotar a individuos, organizaciones o empresas, a menudo engañándolos para que entreguen dinero o información personal. Las estafas comunes incluyen el phishing, donde correos electrónicos y sitios web fraudulentos suplantan la identidad de fuentes legítimas; el fraude de pago por adelantado, que promete recompensas a cambio de pagos anticipados; las estafas románticas que utilizan perfiles falsos de citas en línea para solicitar dinero; y los esquemas de inversión que prometen altos rendimientos. Esto a diferencia de Heiding et al. (2023), quienes centran su análisis en la capacidad de los modelos de lenguaje para generar y detectar mensajes de phishing, Gressel et al. (2026) muestran que la IA también puede intervenir en formas de engaño más prolongadas, como las estafas románticas.

Esta diferencia permite observar que el fraude asistido por inteligencia artificial no opera únicamente mediante mensajes breves, sino también mediante relaciones comunicativas sostenidas en el tiempo. A esto se deme sumar las estafas de fraudes financieros donde se simula operaciones reales de bancos u operadores de tarjetas de créditos para lograr que el usuario entregue su información personal y bancario, para posteriormente proceder a realizar un retiro o transferencia con esta información proporcionada.

A pesar de todo lo expuesto es innegable como la inteligencia artificial se ha convertido en parte integral de las operaciones de las instituciones financieras. La implementación de la IA permitió una mejora significativa en la calidad del servicio y posibilitó soluciones innovadoras para los clientes. Al mismo tiempo, con todas las ventajas y aspectos positivos del uso de la IA, también crea riesgos adicionales, dependiendo de quién y para qué se utilice En el ámbito financiero, la inteligencia artificial puede ser utilizada para simular comunicaciones bancarias, crear perfiles aparentemente legítimos, coordinar interacciones fraudulentas y aumentar la credibilidad de operaciones falsas dirigidas a obtener datos sensibles o transferencias económicas indebidas (Ren et al., 2026; Shpachuk et al., 2026; Vecchietti et al., 2025).

Las estafas de seducción romántica se han convertido en una importante fuente de daño financiero y emocional en todo el mundo, asistidas por inteligencia artificial esta modalidad de estafa muestra que el engaño digital puede desarrollarse de manera progresiva, mediante conversaciones prolongadas que simulan afecto, confianza e intimidad emocional con el propósito de inducir a la víctima a realizar inversiones o transferencias fraudulentas (Gressel et al., 2026; Hossain et al., 2025; Jiang, 2024). Estas operaciones son dirigidas por sindicatos del crimen organizado que trafican con miles de personas para trabajos forzados, obligándolas a construir una relación de intimidad emocional con las víctimas durante semanas de conversaciones por texto antes de presionarlas para que realicen inversiones fraudulentas en criptomonedas. Debido a que las estafas se basan inherentemente

en texto, plantean preguntas urgentes sobre el papel de los modelos de lenguaje a gran escala tanto en la actualidad como en el futuro. (Gressel et al., 2026)

Bajo esta modalidad sucede algo similar a la práctica del phishing, pero con un fuerte elemento de comunicación y construcción de relaciones que incluso puede llevar meses en desarrollarse, lo relevante de esta modalidad es el uso de la IA para lograr que se simule una verdadera conexión real con una persona que se encuentra en situación de vulnerabilidad. Las nuevas modalidades de fraude digital no dependen únicamente de vulnerabilidades técnicas, sino de la explotación de factores humanos como la confianza, la urgencia, el miedo, la autoridad aparente y la falta de verificación del mensaje recibido (Jabir et al., 2025; Matecas et al., 2025; Schmitt & Flechais, 2024).

Los agentes de inteligencia artificial introducen una nueva dimensión en el fraude digital, debido a que pueden simular conversaciones humanas, coordinar respuestas y sostener interacciones prolongadas que aumentan la credibilidad del engaño frente a la víctima (Badhe, 2025; Hossain et al., 2025; Ren et al., 2026). Estas tareas suelen ser gestionadas por varios agentes que trabajan juntos con una división precisa del trabajo. Paralelamente, otra línea de investigación explora las sociedades de agentes, donde a los agentes se les otorga autonomía e interés propio, y las interacciones a gran escala pueden dar lugar a fenómenos sociales emergentes como la cooperación. Estas sociedades pueden utilizarse para estudiar dinámicas sociales complejas y también para simular actividades que implican riesgos éticos. Entre estos riesgos, el fraude financiero es uno de los más perjudiciales.

El rápido crecimiento de las plataformas de redes sociales ha amplificado aún más esta amenaza al proporcionar un terreno fértil para que el fraude se propague a gran escala. (Ren et al., 2026) La interacción entre varios agente puede entregar al usuario una sensación de que se está interactuando con un perfil real, debido a que el sujeto observa que existe interacción dentro de las redes sociales por parte de este usuario ficticio, esta simulación puede llevar a que el usuario sienta la confianza suficiente para entregar información personal, siendo esta situación una de las principales transformaciones introducidas por la inteligencia artificial es la personalización del fraude, pues los modelos generativos pueden adaptar mensajes, perfiles o interacciones a las características particulares de la víctima, aumentando la apariencia de legitimidad y la probabilidad de éxito del engaño (Denisenko et al., 2026; Hazell, 2023; Heiding et al., 2023; Ren et al., 2026).

Los deepfakes (vídeos, audios e imágenes artificiales pero hiperrealistas creados mediante algoritmos) son uno de los últimos avances tecnológicos en inteligencia artificial. Gracias a la velocidad y el alcance de las redes sociales, pueden llegar rápidamente a millones de personas y provocar una amplia gama de engaños en el mercado. Sin embargo, el conocimiento actual sobre las implicaciones de los

deepfakes en el mercado es limitado y fragmentado. (Mustak et al., 2023) Cabe tomar en cuenta que el desarrollo de generación de imágenes y videos de la IA va mejorando cada día, por lo cual es complejo determinar todos los delitos que potencialmente se pueden llevar a cabo por medio del uso de los deepfakes. Existen tres tipos principales de métodos de detección: clásicos, basados en aprendizaje automático y aprendizaje profundo, e híbridos. Existen tres tipos principales de métodos preventivos: técnicos, legales y morales. (Abdel-Wahab & Alkhatib, 2026)

Los deepfakes como resultados de la inteligencia artificial generativa, pueden ser útiles para crear simulaciones realistas en la educación, el periodismo y las artes. Sin embargo, la aparición de estafas maliciosas con deepfakes ha generado preocupación sobre la calidad y la fiabilidad de la información proporcionada a los usuarios de redes sociales (Lui, A and Miglionico 2026) A pesar de esto es innegable el uso doloso que se puede dar a los deepfakes como instrumentos para cometer estafas o incluso extorsiones en el ámbito digital, por lo que para el derecho penal no se discute la veracidad de la información sino la posibilidad de ocasionar un impacto a la vida de esta persona para potencialmente cometer un delito. Así mismo la trata de personas, el lavado de dinero, el chantaje, la pornografía de venganza y el ransomware son algunas de las amenazas asociadas a los deepfakes de las que cualquiera puede ser víctima. Esta es la nueva normalidad de los deepfakes impulsados por inteligencia artificial que no solo pone en cuestión el patrimonio ciudadano, sino que también tiene la posibilidad de afectar otro tipo de cuestiones esenciales para la vida en sociedad (Vecchietti, et al., 2025)

El uso de inteligencia artificial en modalidades como phishing, deepfakes, clonación de voz y agentes conversacionales genera nuevos desafíos probatorios para el derecho penal, especialmente en la identificación del autor, la trazabilidad del engaño, la autenticidad de la evidencia digital y la demostración del nexo entre la conducta engañosa y el perjuicio patrimonial (Gressel et al., 2026; Mustak et al., 2023; Ren et al., 2026; Romero-Moreno, 2025).Los avances en inteligencia artificial generativa han facilitado la suplantación de identidad, permitiendo a los usuarios crear imágenes realistas de personas completamente nuevas o de individuos de confianza. Si bien la susceptibilidad a la inautenticidad basada en mensajes (por ejemplo, el phishing) está bien investigada, aún no está claro si existen mecanismos cognitivos similares que respalden la detección de inautenticidad en técnicas basadas en mensajes e imágenes (por ejemplo, rostros generados por IA) (Sarno et al., 2026)

Los deepfakes han ampliado las posibilidades de suplantación de identidad en entornos digitales, al permitir la creación de imágenes, audios y videos hiperrealistas que pueden ser utilizados para fraudes financieros, manipulación informativa, extorsión o engaños patrimoniales (Mustak et al., 2023; Romero-Moreno, 2025; Vecchietti et al., 2025).Por lo que los deepfakes, habitualmente utilizados para el fraude financiero, son utilizados también en campañas de desinformación política, la

difusión de imágenes no consensuales y el acoso dirigido, además de representar una amenaza en rápida evolución para la integridad de la información global, por su potencial uso masivo exige una intervención inmediata y coordinada por parte las autoridades locales e incluso en muchos casos un esfuerzo y coordinación regional e internacional, esto debido al impacto de estos delitos, dado a que no solo se trata de una afectación que puede darse a nivel individual sino que potencialmente el uso de los deepfakes puede afectar a una sociedad en su conjunto alterando elecciones, afectando el honor de personajes públicos y demás modalidades de afectación. (Romero-Moreno, 2025)

Resultados

Tabla 1. Caracterización de la muestra documental sobre estafa

N.º	Autor/Año	País / afiliación principal	Metodología usada por el autor	Hallazgo principal
1	Schmitt y Flechais (2024)	Reino Unido	Revisión sistemática y propuesta de modelo conceptual	La IA generativa amplifica la ingeniería social mediante creación de contenido realista, personalización avanzada e infraestructura automatizada de ataque.
2	Heiding et al. (2023)	Estados Unidos	Estudio experimental comparativo con participantes y modelos de lenguaje	Los LLM pueden generar correos de phishing personalizados y reducir costos del ataque; además, algunos modelos detectan intención maliciosa con alta capacidad.
3	Hazell (2023)	Reino Unido	Estudio experimental con generación de mensajes para spear phishing	Los LLM pueden apoyar la fase de reconocimiento y la generación de mensajes de spear phishing, incluso a bajo coste.
4	Jabir, Le y Nguyen (2025)	Australia	Revisión sistemática con protocolo PRISMA y metodología Kitchenham	El phishing con IA generativa explota diversos factores humanos como confianza, urgencia, error y falta de alfabetización digital.
5	Hossain et al. (2025)	Estados Unidos	Diseño y evaluación experimental de un framework con LLM y aprendizaje federado	Un sistema AI-in-the-loop puede detectar y ralentizar conversaciones fraudulentas en tiempo real, preservando la privacidad mediante aprendizaje federado.
6	Mai et al. (2026)	Reino Unido	Evaluación multi-turn con expertos en política pública y aplicación de la ley	Los LLM actuales ofrecen ayuda limitada para delitos complejos, pero los modelos sin salvaguardas o con solicitudes fragmentadas pueden aumentar el riesgo de abuso.
7	Ren et al. (2026)	China	Benchmark y simulación de escenarios de fraude con agentes LLM colaborativos	Los agentes LLM pueden coordinarse en fraudes financieros en plataformas sociales, aumentando el riesgo por interacción colectiva y adaptación al entorno.
8	Jiang (2024)	China	Propuesta técnica y evaluación preliminar con GPT-3.5 y GPT-4	Los LLM pueden utilizarse para detectar señales de estafa, phishing, fraude de pago anticipado y romance scams.
9	Gressel et al.	India / Israel /	Entrevistas, estudio	Las estafas románticas pueden

N.º	Autor/Año	País / afiliación principal	Metodología usada por el autor	Hallazgo principal
	(2026)	Italia / Australia	conversacional ciego y evaluación de filtros de seguridad	automatizarse con LLM; los agentes generados por IA pueden producir mayor confianza y cumplimiento que operadores humanos.
10	Shpachuk, Markova y Adamyk (2026)	Reino Unido / Ucrania	Análisis jurídico-regulatorio	El fraude financiero impulsado por IA exige mayor responsabilidad, gestión de riesgos y actualización de mecanismos legales de protección.
11	Chlasta (2025)	Polonia	Análisis teórico-documental sobre ciberseguridad	La IA generativa tiene un carácter dual: puede utilizarse tanto para fortalecer defensas como para facilitar ataques de ingeniería social y phishing.
12	Matecas, Kieseberg y Tjoa (2025)	Austria	Estudio técnico-exploratorio sobre herramientas abiertas de IA	La IA reduce la barrera de entrada para ataques de ingeniería social al permitir generar múltiples plantillas persuasivas en segundos.
13	Denisenko et al. (2026)	Estados Unidos	Estudio aprobado por IRB y análisis estadístico de más de 1.400 hipótesis	La susceptibilidad al phishing depende de la interacción entre características del usuario y propiedades del contenido generado por IA.
14	Lui y Miglionico (2026)	Reino Unido	Análisis jurídico de responsabilidad y protección del consumidor	Las estafas financieras con deepfakes evidencian vacíos de responsabilidad en los marcos de protección del consumidor.
15	Mustak et al. (2023)	Finlandia / Reino Unido / India	Revisión conceptual y análisis de literatura sobre deepfakes	Los deepfakes generan engaños en el mercado, afectan la confianza del consumidor y plantean riesgos para empresas y usuarios.
16	Vecchietti, Liyanaarachchi y Viglia (2025)	Reino Unido / Italia	Estudio cualitativo explicativo con 27 gerentes bancarios de tres bancos globales en nueve países	La integridad de los datos y la gobernanza de IA son claves para mitigar amenazas de deepfakes en el sector bancario.
17	Romero-Moreno (2025)	Reino Unido	Análisis técnico y jurídico-comparado	Los deepfakes usados en fraude, desinformación y acoso requieren detección, trazabilidad, marcas de agua y marcos jurídicos protectores de derechos humanos.
18	Badhe (2025)	No especificado claramente en la fuente consultada	Diseño de agente autónomo y simulación multi-turn de llamadas fraudulentas	Los agentes LLM pueden simular llamadas de estafa realistas, adaptar respuestas y evadir salvaguardas cuando el ataque se fragmenta en varios turnos.
19	Abdel-Wahab y Alkhatib (2026)	Arabia Saudita	Revisión de técnicas de detección y prevención de deepfakes	Las defensas contra deepfakes deben integrar métodos clásicos, machine learning, deep learning, estrategias híbridas y medidas legales o éticas.
20	Sarno et al. (2026)	Estados Unidos	Estudio empírico comparativo sobre detección de inautenticidad digital	Los usuarios detectan peor la inautenticidad basada en imágenes generadas por IA que las estafas basadas en mensajes.

Los resultados de la revisión documental realizada fueron organizados en cinco temáticas, esto de acuerdo con la recurrencia de los problemas abordados por los documentos analizados en este estudio: phishing e ingeniería social asistida por inteligencia artificial, deepfakes y suplantación de identidad, fraudes financieros digitales, estafas románticas automatizadas y carácter dual de la inteligencia artificial.

En primer lugar, se identificó que el phishing, el spear phishing y la ingeniería social constituyen las modalidades más abordadas por la literatura reciente sobre este fenómeno. Los estudios muestran que los modelos de lenguaje permiten generar mensajes fraudulentos más personalizados, coherentes y difíciles de detectar. Esta modalidad se diferencia del phishing tradicional porque la inteligencia artificial puede adaptar el mensaje al perfil de la víctima, utilizar información disponible en redes sociales y simular patrones de comunicación humana para lograr el objetivo delictivo.

En segundo lugar, los deepfakes aparecen como una modalidad emergente que presenta un alto riesgo. La literatura evidencia que la inteligencia artificial permite crear audios, imágenes y videos hiperrealistas que pueden ser utilizados para suplantar identidades, simular comunicaciones legítimas e incluso inducir a la víctima a realizar actos patrimoniales perjudiciales. Esta modalidad genera además dificultades como prueba dentro de un proceso, especialmente en lo que respecta a la autenticidad de la evidencia digital y la identificación del autor del material.

En tercer lugar, se encontraron estudios relacionados con fraudes financieros digitales y agentes de inteligencia artificial. Estos estudios demuestran que la IA puede ser utilizada para simular comunicaciones bancarias, perfiles empresariales, operaciones financieras y escenarios de interacción coordinada en plataformas sociales. La presencia de agentes LLM colaborativos evidencia el riesgo, debido a que el engaño puede adquirir una apariencia de interacción social real.

En cuarto lugar, se identificaron estafas románticas asistidas por modelos de lenguaje. Esta modalidad se caracteriza por la construcción progresiva de vínculos emocionales falsos mediante conversaciones reales, simulación de empatía y generación de confianza personal. A diferencia del phishing tradicional, el engaño no se produce necesariamente en un solo acto, sino a través de una relación sostenida que puede ocasionar que se den inversiones falsas, transferencias o entrega de información sensible.

Finalmente, la revisión permitió identificar el carácter dual de la inteligencia artificial. Por un lado, puede ser empleada como herramienta delictiva para automatizar, personalizar y masificar el engaño. Por otro lado, también puede utilizarse para detectar patrones sospechosos, identificar contenido generado artificialmente y prevenir fraudes digitales. En consecuencia, el problema jurídico no se encuentra en

la inteligencia artificial como tecnología en sí misma, sino en el uso que se le da dentro de contextos de engaño patrimonial.

Discusión

Los resultados de la revisión permiten sostener que la inteligencia artificial no elimina la estructura tradicional de la estafa, sino que transforma los medios mediante los cuales se produce el engaño. En este sentido, el phishing, los deepfakes, las estafas románticas, los fraudes financieros y los agentes conversacionales no constituyen necesariamente delitos completamente nuevos, sino formas actualizadas de engaño patrimonial. Lo que cambia realmente es la capacidad tecnológica para automatizar, personalizar y masificar la conducta fraudulenta a los intereses delictivos.

Esta idea coincide con los aportes de Schmitt y Flechais (2024), Heiding et al. (2023) y Jabir et al. (2025), quienes advierten en sus estudios que la inteligencia artificial generativa permite crear mensajes fraudulentos más creíbles, personalizados y difíciles de detectar. Por lo tanto, la principal aplicación no se encuentra únicamente en la existencia del fraude digital, sino en la facilidad con la que actualmente puede producirse contenido de esta naturaleza a gran escala. Este hecho reduce las barreras técnicas para los ciberdelincuentes y aumenta la exposición de los usuarios a modalidades de estafa más sofisticadas a las habituales.

Asimismo, los hallazgos muestran que el phishing y el spear phishing son las modalidades más desarrolladas en la literatura analizada. Esto se explica debido a que los modelos de lenguaje pueden adaptar mensajes al perfil de la víctima, utilizando información disponible en redes sociales, contextos laborales o patrones de comunicación personal. En este punto, Hazell (2023), Heiding et al. (2023) y Denisenko et al. (2026) coinciden en que la personalización del mensaje incrementa la credibilidad del engaño y el riesgo de exponerse a un fraude digital. Desde una perspectiva jurídico-penal, esto resulta relevante porque el engaño deja de ser genérico y se convierte en una estrategia dirigida, capaz de inducir a error de manera más eficiente que las que habitualmente existían.

Por otra parte, los deepfakes representan un nuevo desafío, ya que afectan la confianza en la imagen, la voz y la identidad digital. Mustak et al. (2023), Romero-Moreno (2025) y Vecchietti et al. (2025) permiten comprender que los deepfakes no solo generan riesgos en el mercado o en la comunicación pública, sino también en el cometimiento de fraudes patrimoniales. La posibilidad de simular la voz de un directivo, la imagen de una persona conocida o una comunicación institucional aparentemente legítima puede facilitar que la víctima se exponga a este tipo de fraude y realice transferencias, entregue información confidencial o confíe en una identidad falsa.

En relación con las estafas románticas y los fraudes financieros, los resultados evidencian que la inteligencia artificial puede sostener interacciones prolongadas con las víctimas. A diferencia del phishing tradicional, que suele operar mediante mensajes breves, las estafas románticas asistidas por modelos de lenguaje pueden construir vínculos emocionales falsos durante días, semanas o incluso meses. Esto coincide con Gressel et al. (2026), quienes muestran que los modelos de lenguaje pueden intervenir en dinámicas de engaño afectivo y financiero. En estos casos, el engaño no se configura en un solo acto, sino mediante una acumulación progresiva de confianza, manipulación y dependencia emocional.

Desde el punto de vista del derecho penal ecuatoriano, estos hallazgos permiten afirmar que las nuevas modalidades de estafa asistida por inteligencia artificial pueden ser analizadas a partir de los elementos tradicionales del delito de estafa: engaño, error, disposición patrimonial, perjuicio y ánimo de lucro. Sin embargo, lo relevante en el campo del derecho es la dificultad probatoria, esto debido a que en los casos donde intervienen deepfakes, agentes conversacionales, perfiles falsos o mensajes generados automáticamente, puede resultar más complejo identificar al autor, demostrar la trazabilidad del engaño y probar la relación entre la conducta fraudulenta y el perjuicio sufrido por la víctima.

También es posible advertir que la respuesta por parte del estado no debe consistir en crear nuevos tipos penales para cada modalidad tecnológica ya que esto ocasiona una excesiva fragmentación normativa podría generar inflación de tipos penales pena y problemas de aplicación práctica, debido a la multiplicidad de tipos. Más bien, el reto consiste en interpretar adecuadamente los tipos penales existentes frente a nuevos medios comisivos, fortalecer las capacidades de investigación digital y mejorar los mecanismos de prevención, trazabilidad y preservación de evidencia electrónica.

Finalmente, los autores mencionados coinciden que la inteligencia artificial debe entenderse como una herramienta de carácter dual en el campo de las estafas digitales. Por un lado, puede ser utilizada por ciberdelincuentes para cometer fraudes más sofisticados, pero, por otro lado, también puede emplearse para detectar patrones de estafa, identificar mensajes sospechosos, analizar conversaciones en tiempo real y prevenir ataques. Por ello, el problema jurídico no radica en la inteligencia artificial en sí misma, sino en su utilización dentro de contextos de engaño patrimonial.

Conclusiones

La revisión documental permitió concluir que la inteligencia artificial ha transformado las modalidades de estafa digital al permitir la automatización, personalización y masificación del engaño que son aplicadas. Herramientas como los modelos de lenguaje, deepfakes, clonación de voz y agentes conversacionales aumentan la credibilidad de las comunicaciones fraudulentas y dificultan su detección por parte

de las víctimas, además del procesamiento de las autoridades. Sin embargo, estas nuevas modalidades no eliminan la estructura clásica de la estafa, sino que modifican sus medios de ejecución: el engaño, el error, la disposición patrimonial, el perjuicio y el ánimo de lucro continúan siendo elementos esenciales para su configuración jurídico-penal, aunque las modalidades sean muy variadas.

Asimismo, se concluye que el phishing, la ingeniería social, los deepfakes, las estafas románticas y los fraudes financieros digitales constituyen las principales manifestaciones de la estafa asistida por inteligencia artificial. El procesamiento de estos delitos por parte de las autoridades por estas modalidades genera retos probatorios desde su etapa de investigación, especialmente en la identificación del autor, la autenticidad de la evidencia digital y la trazabilidad del engaño.

Finalmente, la inteligencia artificial debe entenderse como una herramienta de carácter dual, capaz de facilitar la comisión de fraudes, pero también de contribuir a su detección y prevención. Por ello, la respuesta jurídica debe combinar interpretación penal adecuada, fortalecimiento de capacidades investigativas, alfabetización digital y cooperación institucional.

Referencias Bibliográficas

Abdel-Wahab, A., & Alkhatib, M. (2026). Toward robust deepfake defense: A review of deepfake detection and prevention techniques in images. *Computers, Materials & Continua, 86*(2), 1–34. https://doi.org/10.32604/cmc.2025.070010

Badhe, S. (2025). *ScamAgents: How AI agents can simulate human-level scam calls*. arXiv. https://doi.org/10.48550/arXiv.2508.06457

Chlasta, K. (2025). The dual-use dilemma of generative artificial intelligence in cybersecurity: Navigating the explosive growth in offensive and defensive applications. *Security and Defence Quarterly, 52*(4). https://doi.org/10.35467/sdq/217364

Denisenko, N., La Gatta, V., Postiglione, M., Sola, L., Chen, Y., & Subrahmanian, V. S. (2026). AI-generated phishing: Combining human behavior with post content to assess susceptibility. *ACM Transactions on Internet Technology, 26*(2), Article 24. https://doi.org/10.1145/3799891

Gressel, G., Pankajakshan, R., Rozenfeld, S., Li, L., Franceschini, I., Achuthan, K., & Mirsky, Y. (2026). *Love, lies, and language models: Investigating AI's role in romance-baiting scams*. arXiv.

Hazell, J. (2023). *Spear phishing with large language models*. arXiv. https://doi.org/10.48550/arXiv.2305.06972

Heiding, F., Schneier, B., Vishwanath, A., Bernstein, J., & Park, P. S. (2023). *Devising and detecting phishing: Large language models vs. smaller human models*. arXiv. https://doi.org/10.48550/arXiv.2308.12287

Hossain, I., Puppala, S., Alam, M. J., & Talukder, S. (2025, November 5). *AI-in-the-loop: Privacy preserving real-time scam detection and conversational scambaiting by leveraging LLMs and federated learning*. arXiv.

Jabir, R., Le, J., & Nguyen, C. (2025). Phishing attacks in the age of generative artificial intelligence: A systematic review of human factors. *AI, 6*(8), Article 174. https://doi.org/10.3390/ai6080174

Jiang, L. (2024). *Detecting scams using large language models*. arXiv.

Lui, A., & Miglionico, A. (2026). AI-generated deepfake financial scams: A missing liability regime for consumer protection frameworks. *Asian Journal of Comparative Law*. Advance online publication.

Mai, K. T., Gausen, A., Dubois, M., Murad, M., O'Dell, B., Staes-Polet, N., Summerfield, C., & Strait, A. (2026, February). *A multi-turn framework for evaluating AI misuse in fraud and cybercrime scenarios*. arXiv.

Matecas, A.-R., Kieseberg, P., & Tjoa, S. (2025). Social engineering with AI. *Future Internet, 17*(11), Article 515. https://doi.org/10.3390/fi17110515

Mustak, M., Salminen, J., Mäntymäki, M., Rahman, A., & Dwivedi, Y. K. (2023). Deepfakes: Deceptions, mitigations, and opportunities. *Journal of Business Research, 154*, Article 113368. https://doi.org/10.1016/j.jbusres.2022.113368

Ren, Q., Zheng, Z., Guo, J., Yan, J., Ma, L., & Shao, J. (2026, April 6). *When AI agents collude online: Financial fraud risks by collaborative LLM agents on social platforms*. arXiv.

Romero-Moreno, F. (2025). Deepfake detection in generative AI: A legal framework proposal to protect human rights. *Computer Law & Security Review, 58*, Article 106162. https://doi.org/10.1016/j.clsr.2025.106162

Sarno, D. M., Solorio, J., Ballar, S., Chadwick, S., Harris, K., Moss, D., & Lyu, S. (2026). Framing digital inauthenticity: Comparing user detection of AI-generated faces to message-based scam methods. *Acta Psychologica, 262*, Article 105995. https://doi.org/10.1016/j.actpsy.2025.105995

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

Shpachuk, V., Markova, O., & Adamyk, B. (2026). AI-driven financial fraud: Key risks and legal protections for financial institutions. *Journal of Banking Regulation, 27*, Article 6. https://doi.org/10.1057/s41261-025-00304-y

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