EL CASO ARUP Y EL FRAUDE MEDIANTE DEEPFAKES EN TIEMPO REAL

Análisis jurídico-técnico de la suplantación de identidad mediante inteligencia artificial generativa y propuestas de reforma normativa

Informe de Investigación Jurídica

Marzo de 2026

ÍNDICE

CAPÍTULO 1. INTRODUCCIÓN: EL CAMBIO DE PARADIGMA EN EL FRAUDE FINANCIERO

1.1. Objeto y justificación del informe

A principios de 2024, la filial hongkonesa de la firma de ingeniería Arup fue víctima de un fraude sin precedentes en la historia de la ciberdelincuencia corporativa. Lo que comenzó como una aparente videoconferencia ordinaria derivó en la transferencia fraudulenta de 25,6 millones de dólares estadounidenses mediante quince transacciones dirigidas a cuentas bancarias locales en Hong Kong.[1] Los autores del delito no emplearon un simple correo electrónico de *phishing* con deficiencias ortográficas, sino que desplegaron algoritmos de inteligencia artificial generativa para suplantar, con un grado de realismo técnicamente extraordinario, al director financiero del grupo radicado en el Reino Unido y a otros directivos de la organización. El fraude no fue descubierto hasta seis días después, cuando el empleado afectado consultó directamente con la sede central de la multinacional, momento en el que los fondos ya habían sido disipados a través de la red bancaria.

La relevancia sistémica de este incidente —catalogado en la *AI Incident Database* como el Incidente 634— reside en que ilustra con precisión el fenómeno que los organismos internacionales denominan el *multiplicador de fuerzas* de la criminalidad organizada habilitado por la inteligencia artificial generativa.[2] Las proyecciones más recientes estiman que las pérdidas derivadas de fraudes impulsados por inteligencia artificial en los Estados Unidos podrían alcanzar los 40.000 millones de dólares en 2027,[3] lo que exige un análisis riguroso de la capacidad del ordenamiento jurídico

[1]AI Incident Database, Incident 634 / Report 3642: Arup deepfake fraud, disponible en: https://incidentdatabase.ai [consulta: marzo 2026].
[2]EL PACCTO 2.0 (Expertise France / FIAP), "Weaponizing Artificial Intelligence: How AI Reshapes the World of Organized Crime", 2025, p. 14.
[3]Deloitte Center for Financial Services, "The coming wave of AI-enabled financial fraud", 2024, estimación proyectada para EE. UU. al horizonte 2027.

vigente para responder a una amenaza que desborda la lógica tradicional del delito patrimonial. El presente informe se propone, en consecuencia, evaluar si el marco normativo actual —incluido el Reglamento (UE) 2024/1689, conocido como *AI Act* o Reglamento de Inteligencia Artificial (RIA)— resulta suficiente o si, por el contrario, se impone una reforma legislativa de calado que tipifique de forma autónoma las conductas delictivas facilitadas por *deepfakes* maliciosos.

1.2. Objetivos de la investigación

El objetivo central de este trabajo consiste en desentrañar el alcance doctrinal y práctico de la responsabilidad penal en casos en los que el agente delictivo opera mediante *algoritmos opacos* e infraestructuras distribuidas. La investigación no se limita a la reconstrucción del incidente, sino que persigue determinar si el ordenamiento jurídico vigente es capaz de responder a una amenaza que supera la lógica tradicional del delito patrimonial. En particular, se evalúa la efectividad de las obligaciones de transparencia del RIA —concretamente el etiquetado de contenidos sintéticos previsto en su artículo 50— frente a la sofisticación de herramientas diseñadas expresamente para el fraude, como la plataforma *Haotian AI*. Asimismo, se pretende verificar si existe una desprotección estructural derivada de la estrechez de unos tipos penales que han quedado desfasados frente a las denominadas *ultrasuplantaciones*. La pregunta central que articula este análisis es si resulta suficiente la aplicación de la estafa informática tradicional o si, por el contrario, se requieren tipos penales autónomos que protejan la integridad digital y el derecho a la propia imagen como bienes jurídicos independientes.[4]

Un tercer objetivo, de naturaleza propositiva, consiste en formular estrategias de ciberresiliencia corporativa que integren la filosofía de *Zero Trust* —confianza cero— en la gestión de transacciones financieras

[4]ISO/IEC 30107-3:2023, Information technology — Biometric presentation attack detection — Part 3: Testing and reporting, International Organization for Standardization, Ginebra, 2023.

remotas de alto valor, asegurando que la autenticación sea continua y no descanse en una verificación exclusivamente visual o auditiva.

1.3. Metodología multidisciplinar

La naturaleza poliédrica del fenómeno exige un diseño metodológico que combine la hermenéutica jurídica con el análisis forense informático y la gestión de riesgos corporativos. La comprensión cabal de la estafa perpetrada contra Arup requiere, inevitablemente, sumergirse en la mecánica de las Redes Adversarias Generativas (GANs) e identificar cómo el proceso de competición entre la red generadora y la red discriminadora permite alcanzar el grado de realismo suficiente para superar los sensores de detección de vida regulados en la norma técnica ISO/IEC 30107-3:2023.[5] El análisis documental se fundamenta en fuentes de alta fiabilidad, entre las que destacan los estudios de Europol[6] y de la iniciativa EL PACCTO 2.0,[7] además de la normativa técnica internacional y la doctrina penal especializada.

La investigación sigue un proceso deductivo: desde la caracterización global de las amenazas de la inteligencia artificial generativa hasta la concreción de propuestas de *lege ferenda*. El objetivo final es que el presente informe no constituya un mero ejercicio académico, sino una herramienta de *seguridad por diseño* que permita a los responsables corporativos y a los operadores jurídicos adoptar decisiones fundadas en un entorno de elevada incertidumbre tecnológica.

[6]EUROPOL, "ChatGPT — The impact of Large Language Models on Law Enforcement", EC3 SPOTLIGHT Report, La Haya, 2023, p. 8.

CAPÍTULO 2. MARCO TÉCNICO: INTELIGENCIA ARTIFICIAL GENERATIVA Y ULTRAFALSIFICACIONES

2.1. Funcionamiento de las Redes Adversarias Generativas (GANs)

Para comprender la magnitud del engaño sufrido por Arup, es preciso examinar la arquitectura técnica que lo hizo posible. Las Redes Adversarias Generativas (GANs), conceptualizadas por Goodfellow *et al.* en 2014,[8] no constituyen un sistema estático de procesamiento, sino un marco de competición entre dos arquitecturas de redes neuronales con objetivos opuestos. La red generadora tiene por misión crear contenido —imágenes, secuencias de vídeo o audio— que resulte indistinguible de la realidad; la red discriminadora actúa, en contrapartida, como mecanismo de control de calidad que evalúa continuamente si el artefacto generado es auténtico o sintético, tomando como referencia los datos de entrenamiento. El proceso de retroalimentación entre ambas redes produce una mejora iterativa que, en términos de teoría de juegos, tiende hacia un equilibrio de Nash:[9] el punto en el que el contenido sintético es técnicamente indistinguible de la realidad para los sistemas de detección convencionales.

La opacidad inherente a estos modelos plantea un desafío jurídico de primer orden: la imposibilidad de auditar el proceso de aprendizaje dificulta la trazabilidad del engaño y, con ello, la atribución de responsabilidad penal a los distintos actores de la cadena delictiva.[10] Esta circunstancia ha llevado a la doctrina a hablar de una *ambigüedad epistémica* inherente a los sistemas de IA generativa, cuyas implicaciones jurídicas se analizan en el Capítulo 4.

---

[8]GOODFELLOW, Ian et al., "Generative Adversarial Networks", Advances in Neural Information Processing Systems, vol. 27 (2014), pp. 2672-2680.
[9]NASH, John F., "Equilibrium Points in N-Person Games", Proceedings of the National Academy of Sciences, vol. 36, núm. 1 (1950), pp. 48-49.
[10]EL PACCTO 2.0, op. cit., p. 22.

2.2. Modalidades técnicas empleadas en el caso Arup

El ataque perpetrado contra Arup no constituyó un *deepfake* genérico, sino un *ataque de presentación* en tiempo real de sofisticación *prima facie* inesperada para los sistemas de verificación corporativos.[11] La mecánica empleada integró dos vertientes técnicas complementarias: la clonación de voz (*voice cloning*) y el intercambio de rostros (*face-swapping*) aplicado a un flujo de vídeo en directo. Lo que el empleado percibió en su pantalla no fue un vídeo pregrabado estático, sino avatares sintéticos que empleaban técnicas de sincronización labial (*lip-sync*) para responder a la interacción en tiempo real. Investigaciones recientes acreditan que, con apenas veinte segundos de audio capturado de una intervención pública disponible en plataformas como YouTube o LinkedIn, es técnicamente factible replicar el timbre, la cadencia y la inflexión emocional de un alto directivo.[12]

En el incidente que nos ocupa, los autores descargaron registros audiovisuales públicos de los directivos suplantados para alimentar el modelo generativo y, posteriormente, doblaron el audio mediante técnicas de IA para integrarlo en la videoconferencia fraudulenta.[13] Esta modalidad técnica genera lo que la doctrina especializada denomina una *ambigüedad epistémica*: la víctima, incluso siendo un profesional experimentado, se ve en la imposibilidad de distinguir la fuente original de la manipulación digital, en particular cuando el ataque se desarrolla en un entorno de alta presión y estricta confidencialidad.

2.3. Herramientas y democratización del fraude

El factor que ha alterado de forma estructural las condiciones de la ciberdelincuencia financiera es la democratización de estas herramientas.

[11]RIBAS, Xavier, "Ataques de presentación mediante deepfakes en videoconferencia", Ribas Artículos, 2024, disponible en: https://ribas.eu [consulta: marzo 2026].
[12]WANG, Yuxuan et al., "Neural Voice Cloning with a Few Samples", Advances in Neural Information Processing Systems, vol. 31 (2018).
[13]AI Incident Database, op. cit., Incident 634, descripción técnica del vector de ataque.

El mercado del cibercrimen ha evolucionado hacia un modelo de *Crime-as-a-Service* (CaaS), en cuyo marco plataformas como *Haotian AI* se comercializan abiertamente en entornos digitales al margen de los mercados regulados.[14] Este software, diseñado expresamente para el fraude, permite realizar intercambios de rostros en tiempo real y clonación de voz sin conocimientos técnicos avanzados, con precios que oscilan entre los 1.200 y los 9.900 dólares.[15]

Herramientas de uso legítimo, como ElevenLabs o el modelo VALL-E de Microsoft —que replica una voz a partir de tan solo tres segundos de muestra de audio—, han sido reconfiguradas para actividades ilícitas, eliminando las barreras de entrada que históricamente limitaban este tipo de ataques a actores estatales o grupos con recursos extraordinarios. La facilidad de acceso a estas tecnologías eleva exponencialmente el nivel de riesgo para los derechos fundamentales reconocidos en la Carta de los Derechos Fundamentales de la Unión Europea (CDFUE), en particular el derecho a la protección de datos personales (art. 8 CDFUE) y el derecho a la propiedad (art. 17 CDFUE), tal como ha sido señalado por la doctrina especializada en el análisis de las amenazas de la IA generativa.[16]

[14]EL PACCTO 2.0, op. cit., p. 31.
[15]Haotian AI pricing structure, documentado por Group-IB Threat Intelligence, "Hi-Tech Crime Trends 2024", Singapur, 2024, p. 47.

CAPÍTULO 3. ANATOMÍA DEL INCIDENTE: LA ESTAFA DE 25,6 MILLONES DE DÓLARES

3.1. Crónica del ataque

La secuencia delictiva se inició en enero de 2024 en la filial hongkonesa de Arup, cuando un empleado del departamento financiero recibió un correo electrónico de *phishing* aparentemente remitido por el director financiero del grupo desde el Reino Unido.[17] En dicha comunicación se le instaba a participar en una videoconferencia confidencial para tratar una supuesta adquisición estratégica. Aunque el empleado experimentó dudas razonables ante la mención de transacciones de carácter secreto, su escepticismo inicial se disipó al incorporarse a la reunión virtual y percibir los rostros y voces de su superior jerárquico y de otros colegas conocidos en la pantalla.

Lo que el trabajador no pudo determinar fue que se encontraba en un entorno de reunión digital en el que él constituía el único participante humano real, rodeado de avatares sintéticos.[18] Los atacantes no generaron vídeos desde cero en tiempo real, sino que descargaron metraje público de conferencias previas de los directivos —disponible en plataformas como YouTube o LinkedIn— y aplicaron algoritmos de IA para doblar el audio y sincronizar los movimientos labiales con una precisión técnica que eliminó los principales artefactos visuales habitualmente asociados a los *deepfakes* de menor calidad, como parpadeos erráticos o inconsistencias en los bordes del rostro.

3.2. Ingeniería social y manipulación psicológica

El éxito de la operación no dependió exclusivamente de la calidad técnica de los avatares sintéticos, sino de una explotación sistemática de las

[17]AI Incident Database, op. cit., Incident 634, cronología del ataque.

vulnerabilidades cognitivas y de las jerarquías corporativas de la víctima.[19] Los autores aplicaron una presión psicológica sostenida, subrayando la urgencia y el carácter reservado de la operación para inhibir cualquier intento de verificación por canales secundarios. La percepción de que varios ejecutivos de la organización respaldaban conjuntamente la instrucción generó en el empleado lo que la psicología social denomina *validación social sintética*: un mecanismo de refuerzo grupal artificial que neutralizó su escepticismo profesional.

Un elemento táctico de particular relevancia fue la estrategia de minimizar la interacción directa con la víctima durante la videoconferencia, limitándose a solicitar una breve presentación inicial, lo que redujo las posibilidades de que las respuestas de los avatares generasen inconsistencias detectables.[20] Los autores reforzaron además la verosimilitud del engaño mediante comunicaciones complementarias por WhatsApp y correo electrónico, articulando un ecosistema de comunicación multicanal que blindaba la narrativa fraudulenta. Desde la perspectiva de la seguridad de la información, este nivel de orquestación evidencia que la *suplantación de identidad digital* (*identity spoofing*) ha alcanzado en el contexto de la IA generativa un grado de sofisticación que desborda los protocolos de cumplimiento normativo anteriores a la irrupción de estas tecnologías.

3.3. Ejecución financiera y revelación del fraude

Bajo la convicción errónea de estar ejecutando una orden directa de la cúpula financiera de Londres, el empleado autorizó el envío de fondos. En total, se ejecutaron quince transferencias bancarias que acumularon la cifra de 200 millones de dólares de Hong Kong (aproximadamente 25,6 millones de dólares estadounidenses). Los fondos fueron dirigidos a cinco

[19]INTERPOL, "Financial Crimes Threat Assessment 2024", INTERPOL General Secretariat, Lyon, 2024, p. 19.
[20]ACALE SÁNCHEZ, María & BOZA MARTÍNEZ, Diego, "La imagen ante el Derecho penal: nuevos desafíos", e-Eguzkilore, Revista del Instituto Vasco de Criminología, núm. 18 (2025), pp. 3-22.

cuentas bancarias locales, elección que facilitó la rapidez de la liquidación antes de que los sistemas de prevención del blanqueo de capitales pudiesen activar las alarmas pertinentes.[21]

El descubrimiento del fraude tuvo lugar seis días después de la videoconferencia, cuando el empleado contactó directamente con la sede central de la multinacional por un trámite contable ordinario. Para entonces, los fondos habían sido transferidos y disueltos en la red bancaria, lo que frustró cualquier posibilidad de recuperación inmediata.[22] Este incidente ha forzado a los reguladores y a organismos como INTERPOL a reconocer que los derechos fundamentales a la integridad patrimonial y a la identidad digital se encuentran sometidos a una amenaza que ha dejado de ser teórica para manifestarse con consecuencias patrimoniales de decenas de millones de dólares.

CAPÍTULO 4. ANÁLISIS JURÍDICO Y REGULATORIO: RESPONSABILIDAD PENAL Y EL REGLAMENTO DE INTELIGENCIA ARTIFICIAL

4.1. Calificación jurídica: estafa informática y falsedad documental

La primera aproximación analítica al caso Arup desde la perspectiva del Derecho penal pone de manifiesto una tensión estructural entre la suficiencia descriptiva de los tipos penales existentes y la especificidad técnica de la conducta enjuiciada. Desde un enfoque técnico-jurídico, nos encontramos ante una estafa informática en la que el engaño se articula mediante una manipulación tecnológica encaminada a inducir en la víctima un acto de disposición patrimonial. Sin embargo, la singularidad del caso reside en que el ardid no consiste en la alteración de un sistema informático, sino en la suplantación de la voluntad humana a través de algoritmos generativos opacos.

La calificación adicional como falsedad documental exige determinar previamente si el flujo de una videoconferencia puede ser considerado documento a los efectos del artículo 26 del Código Penal español. La doctrina más reciente y la jurisprudencia del Tribunal Supremo se inclinan favorablemente por dicha calificación, entendiendo que todo soporte informático que incorpore datos o declaraciones con eficacia probatoria constituye, a efectos legales, un documento digital.[23] Si se acepta esta premisa —como resulta jurídicamente razonable—, el empleo de un *deepfake* para autorizar transferencias de 25,6 millones de dólares no solo integra la estafa informática del artículo 248.2 CP, sino que entra de lleno en el campo de la falsedad en documento privado o incluso mercantil. La alteración del rostro y la voz de los ejecutivos suplantados constituye una falsificación de la verdad en un soporte que la víctima tuvo por auténtico,

[23]STS (Sala de lo Penal) de 21 de octubre de 2021, recurso núm. 1167/2020, sobre la naturaleza del soporte informático como documento a efectos del art. 26 CP.

elemento nuclear del tipo del artículo 395 CP en relación con el artículo 390.1.4.º CP.[24]

El problema que subsiste, con todo, es el de la atribución de autoría. La arquitectura técnica de los *deepfakes* dispersa la responsabilidad a través de una cadena de servidores remotos y jurisdicciones múltiples, dificultando la individualización de los intervinientes y la concreción de los títulos de imputación, cuestión que ha sido analizada en perspectiva comparada por Sánchez Salazar *et al.* en el contexto iberoamericano.[25]

4.2. El Reglamento (UE) 2024/1689 y los desafíos de la transparencia

En el marco regulatorio europeo, el Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024[26] —cuya entrada en vigor plena se produce de forma escalonada hasta agosto de 2026— establece un régimen de obligaciones diferenciadas en función del nivel de riesgo del sistema de inteligencia artificial. El artículo 50.2 RIA impone al operador de un sistema de IA que genere o manipule contenido de imagen, audio o vídeo de forma que produzca una semejanza apreciable con personas u objetos reales existentes, la obligación de revelar que dicho contenido ha sido generado o manipulado artificialmente.[27] El artículo 50.4 RIA extiende esta obligación a los proveedores de sistemas de IA de uso general cuando estos sean capaces de generar contenido sintético de naturaleza visual, auditiva o textual.[28]

No obstante, la eficacia preventiva de estas obligaciones de transparencia frente al crimen organizado plantea interrogantes de primer orden. La

[24]Convención de Budapest sobre la Ciberdelincuencia, CETS núm. 185, Consejo de Europa, 2001, artículos 7 y 8, aplicables a la falsedad y el fraude informáticos.
[25]SÁNCHEZ SALAZAR, P. M. et al., "Responsabilidad penal por manipulación digital con fines delictivos", Revista Cuestiones Políticas, vol. 40, núm. 73 (2022), pp. 212-235.
[26]Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial [en adelante, RIA o AI Act], DO L, 12 de julio de 2024. Considerando 133.
[27]Art. 50.2 RIA.
[28]Art. 50.4 RIA. Véase también ESTELLA, A., "Regulación de las 'ultrasuplantaciones' en el Reglamento de la UE sobre IA", Revista de Administración Pública, núm. 221 (2023), pp. 89-120.

obligación de etiquetado carece de toda efectividad disuasoria cuando el emisor es precisamente una red criminal cuyo objetivo estratégico consiste en suprimir cualquier marca de autenticidad. Ello evidencia una limitación estructural del RIA en su vertiente preventiva: la norma está concebida fundamentalmente para regular a los operadores legítimos del mercado, no para neutralizar el uso delictivo de herramientas de IA.[29]

En lo que respecta a la clasificación del riesgo, los sistemas capaces de generar *deepfakes* de alta resolución en tiempo real podrían quedar comprendidos en el Anexo III, punto 6, del RIA cuando se utilicen en ámbitos que afecten a la seguridad crítica o a los derechos fundamentales de las personas,[30] lo que impondría a sus proveedores y responsables del despliegue un conjunto exigente de obligaciones previas a la comercialización, incluyendo la evaluación de conformidad y el registro en la base de datos de la UE. El incidente de Arup demuestra, sin embargo, que la IA generativa no solo constituye una herramienta de desinformación política, sino un multiplicador de fuerzas para el fraude financiero a gran escala, lo que justifica una interpretación extensiva de los supuestos de alto riesgo.

Un vacío normativo de especial relevancia se concreta en la retirada de la propuesta de Directiva sobre responsabilidad civil en materia de inteligencia artificial por parte de la Comisión Europea en su programa de trabajo para 2025.[31] Sin un régimen armonizado de responsabilidad civil y penal que determine con precisión quién responde por los daños ocasionados por un algoritmo autónomo de aprendizaje, la incertidumbre jurídica resultante opera en beneficio del infractor.

Desde la perspectiva del derecho comparado, el ordenamiento de los Estados Unidos aborda parcialmente esta cuestión a través de la *Executive*

[30]Art. 6 en relación con el Anexo III, punto 6, RIA, versión consolidada DO L 2024.
[31]Propuesta de Directiva del Parlamento Europeo y del Consejo relativa a la responsabilidad civil extracontractual por daños causados por sistemas de inteligencia artificial, COM(2022) 496 final, retirada formalmente en el programa de trabajo de la Comisión para 2025.

Order on Safe, Secure, and Trustworthy AI (EO 14110, octubre de 2023),[32] que impone obligaciones de revelación de contenido sintético a los desarrolladores de modelos de fundación. En el ámbito asiático, la República Popular China ha promulgado las *Regulations on the Management of Deep Synthesis Technology*, en vigor desde el 10 de enero de 2023, que obligan a los proveedores de tecnología de síntesis profunda a marcar el contenido generado y a verificar la identidad real de los usuarios.[33] En el plano internacional, el Convenio Marco sobre Inteligencia Artificial y Derechos Humanos del Consejo de Europa (CETS núm. 225, 2024) constituye el primer tratado internacional jurídicamente vinculante en materia de IA, e impone a los Estados parte obligaciones de protección frente a los riesgos que estos sistemas generan para los derechos fundamentales.[34]

4.3. El vacío legal y la necesidad de tipos penales autónomos

La investigación conduce a una conclusión de fondo: el sistema penal vigente es reactivo y estructuralmente fragmentado frente al fenómeno de las *ultrasuplantaciones*. En el caso Arup, la lesión no se limitó a la dimensión patrimonial del fraude, sino que comprendió una vulneración flagrante de la identidad digital y del derecho a la propia imagen de los ejecutivos suplantados, bienes jurídicos que la doctrina penal clásica ha tendido a considerar como intereses secundarios o instrumentales. Sin embargo, ante la generalización de las ultrafalsificaciones, cabe sostener que la integridad de la imagen digital debe ser protegida como bien jurídico autónomo, con independencia de que su lesión venga acompañada de un perjuicio patrimonial.[35]

[32]Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence (EO 14110), 30 de octubre de 2023, Sección 4.2, sobre obligaciones de revelación de contenido sintético.
[33]Regulations on the Management of Deep Synthesis Technology, Administración del Ciberespacio de China, en vigor desde el 10 de enero de 2023, artículo 6.
[34]Convenio Marco sobre Inteligencia Artificial y Derechos Humanos, Democracia y Estado de Derecho, CETS núm. 225, Consejo de Europa, abierto a la firma el 5 de septiembre de 2024.

En el contexto jurídico español, los tribunales han debido recurrir al delito de vejaciones o trato degradante y a la protección de la integridad moral para colmar el vacío que deja la ausencia de una tipificación específica del *deepfake* malicioso.[36] El Proyecto de Ley Orgánica de protección de las personas menores de edad en los entornos digitales ya contempla una aproximación en esta dirección,[37] pero su ámbito de aplicación se circunscribe a los menores y no da respuesta al fraude económico a gran escala. Esta elasticidad de los tipos penales existentes, aunque útil a corto plazo, genera una inseguridad jurídica incompatible con las exigencias del principio de legalidad penal (*lex certa*), que requiere que las conductas delictivas estén descritas con la suficiente precisión para que el ciudadano pueda prever las consecuencias jurídicas de sus actos.

Resulta, por tanto, imperativo que las reformas de *lege ferenda* protejan la dignidad humana y la veracidad de las interacciones digitales frente a una tecnología capaz de socavar la confianza mínima necesaria para el funcionamiento de las relaciones comerciales y sociales. Las propuestas concretas de reforma se desarrollan en el Capítulo 6.

[36]ACALE SÁNCHEZ & BOZA MARTÍNEZ, op. cit., pp. 15-18.
[37]Proyecto de Ley Orgánica de protección de las personas menores de edad en los entornos digitales, BOCG, Congreso de los Diputados, Serie A, núm. 51-1, XIV Legislatura (2023).

CAPÍTULO 5. ESTRATEGIA CORPORATIVA Y GESTIÓN DEL RIESGO: HACIA LA CIBERRESILIENCIA

5.1. Implementación del SGSI (ISO/IEC 27001): el compromiso de la dirección

El desastre financiero de Arup no constituyó exclusivamente un fallo técnico, sino una quiebra de gobernanza de la información que evidencia la necesidad de orientar las inversiones en seguridad hacia la creación de una cultura de resiliencia organizativa. La ciberseguridad en la era de la inteligencia artificial generativa ha dejado de ser responsabilidad exclusiva del departamento de sistemas para convertirse en una disciplina transversal que requiere el compromiso explícito e inequívoco de la alta dirección.[38] En este marco, la implantación de un Sistema de Gestión de Seguridad de la Información (SGSI) estructurado bajo el estándar ISO/IEC 27001:2022[39] permite alinear la estrategia de seguridad con los objetivos de negocio y establecer un marco normativo interno de aplicación transversal.

La ciberresiliencia debe ser entendida no como la eliminación del riesgo de ataque —objetivo que resulta inalcanzable en el contexto tecnológico actual—, sino como la capacidad operativa para mantener las funciones esenciales de la organización y recuperarse con celeridad tras la materialización de una brecha de seguridad. Esta distinción conceptual tiene implicaciones directas sobre la asignación de recursos y sobre el diseño de los protocolos de respuesta a incidentes.

5.2. Protocolos de verificación frente a IA y la filosofía Zero Trust

[38]CEIM (Confederación Empresarial de Madrid-CEOE), "Informe sobre ciberseguridad y su impacto en las empresas", 2025, pp. 22-30.
[39]ISO/IEC 27001:2022, Information security, cybersecurity and privacy protection — Information security management systems — Requirements, International Organization for Standardization, Ginebra, 2022.

El incidente de Arup ha evidenciado la obsolescencia de los protocolos de verificación que descansan exclusivamente en la comprobación visual o auditiva de la identidad en entornos remotos. En un contexto en el que los actores maliciosos aplican ataques de presentación en tiempo real, la filosofía de *Zero Trust* —concebida originalmente como principio de arquitectura de sistemas— debe trascender su formulación técnica para convertirse en una norma operativa de gestión de transacciones financieras de alto valor: ningún usuario, dispositivo ni instrucción puede considerarse auténtico por el mero hecho de presentar una apariencia familiar, y toda transacción significativa debe ser confirmada por una vía secundaria e independiente.

Como ilustración práctica de la eficacia de esta filosofía, cabe citar el caso de Ferrari, en el que un ejecutivo de la compañía, ante la sospecha de estar interactuando con un interlocutor artificialmente generado, formuló una pregunta personal que solo el verdadero director ejecutivo podría responder, logrando así que el estafador interrumpiese la comunicación de inmediato.[40] Esta prueba de vida contextual ejemplifica el tipo de mecanismos de verificación fuera de guion que las organizaciones deben integrar en sus protocolos de autorización de transferencias de alto valor.

En el plano tecnológico, resulta asimismo fundamental incorporar soluciones de detección de ataques de presentación (*liveness detection*) que identifiquen inconsistencias biológicas —parpadeos irregulares, reflejos antinaturales en el iris, anomalías en la sincronización labial— conforme a los estándares de prueba de la ISO/IEC 30107-3:2023.[41] Estas tecnologías actúan como un estrato adicional de verificación que complementa los protocolos organizativos y reduce la dependencia de la percepción humana como único mecanismo de autenticación.

[40]Cybersecurity Law Report, op. cit., pp. 14-15 (caso Ferrari).
[41]ISO/IEC 30107-3:2023, op. cit., sección 6.

5.3. Formación, concienciación y defensa frente al vibe hacking

El empleado constituye, en última instancia, la primera y la última línea de defensa frente a la ingeniería social potenciada por inteligencia artificial. Los programas de formación en ciberseguridad de generaciones anteriores han quedado insuficientes ante la emergencia del *vibe hacking*: la técnica mediante la cual la IA no solo replica el rostro o la voz de un directivo, sino su estilo comunicativo, su cadencia y su urgencia emocional característica, con el propósito de neutralizar el juicio crítico del destinatario del engaño.[42]

Los programas de capacitación deben incorporar simulaciones prácticas de ataques de *deepfake* adaptadas a los perfiles de riesgo de los equipos financieros y administrativos, con el objetivo de desarrollar un escepticismo profesional sistematizado frente a toda solicitud que se desvíe de los procesos ordinarios de autorización.[43] La seguridad por diseño debe ser, en consecuencia, una convergencia entre la pericia técnica y una cultura organizacional que anteponga la cautela a la celeridad imprudente, garantizando que la tecnología opere como un instrumento de progreso y no como un vector de exposición patrimonial.

CAPÍTULO 6. CONCLUSIONES Y PROPUESTAS DE FUTURO

6.1. Impacto sistémico del caso Arup

El denominado Incidente 634 ha trascendido su condición de episodio aislado de ciberdelincuencia para convertirse en el hecho de referencia ineludible en el estudio del fraude financiero de nueva generación.[44] La estafa de 25,6 millones de dólares perpetrada contra la filial hongkonesa de Arup acredita que hemos entrado en una era en la que la percepción visual y auditiva ha dejado de constituir un ancla fiable para la verificación de la identidad en entornos remotos. El ataque de presentación multizona documentado en este incidente —que no se limitó a clonar una voz, sino que orquestó una reunión virtual íntegramente poblada por avatares sintéticos— revela la profundidad del cambio de paradigma que la IA generativa ha introducido en la arquitectura del riesgo corporativo.

Las proyecciones que estiman pérdidas por fraudes impulsados por IA de hasta 40.000 millones de dólares en 2027[45] señalan que la vulnerabilidad detectada en Hong Kong no es un fenómeno local, sino la manifestación anticipada de una amenaza sistémica global. La principal conclusión operativa es que la ciberresiliencia ya no puede limitarse a la protección de los servidores e infraestructuras tecnológicas, sino que debe extenderse a la verificación continua de la autenticidad de las interacciones humanas en entornos digitales, dotando de contenido operativo real a los principios de la filosofía *Zero Trust*.

6.2. Recomendaciones de lege ferenda

Desde una perspectiva estrictamente jurídica, el marco normativo vigente se revela insuficiente. Resulta imperativo que el legislador —tanto en el ámbito de la Unión Europea como en el de los ordenamientos nacionales— abandone la concepción del *deepfake* como mero agravante de la estafa o

como instrumento de desinformación política, para reconocerlo como un fenómeno autónomo que lesiona bienes jurídicos específicos que merecen protección independiente. En concreto, se formulan las siguientes propuestas de *lege ferenda*:

Primera: creación de un tipo penal autónomo que criminalice la generación o difusión de *deepfakes* maliciosos cuando tengan por finalidad la defraudación patrimonial o la lesión de la integridad moral de la persona suplantada, con independencia de que se consume el resultado dañoso.[46] Segunda: modificación del Reglamento (UE) 2024/1689 para incorporar un régimen de responsabilidad objetiva que obligue a los proveedores de sistemas de IA generativa capaces de producir *deepfakes* de alta resolución a contribuir a los mecanismos de reparación de los daños causados por el uso malicioso de dichos sistemas.[47] Tercera: transposición anticipada de las disposiciones del RIA relativas a las obligaciones de transparencia, con plazos de cumplimiento acelerados para los sectores financiero y de infraestructuras críticas, de conformidad con el calendario establecido en el artículo 113 RIA.

En el contexto español, la incorporación de estas reformas exige una coordinación entre la legislación penal especial, las normas de transposición del RIA y el Proyecto de Ley Orgánica de protección de menores en entornos digitales,[48] evitando la fragmentación normativa que actualmente genera incertidumbre sobre el título de imputación aplicable en casos de fraude mediante *deepfakes*.

6.3. Visión estratégica 2025-2027: inteligencia compartida y marcos de cooperación

El futuro de la lucha contra el cibercrimen potenciado por inteligencia artificial no reside en el aislamiento institucional, sino en una cooperación público-privada estructurada que hoy se encuentra todavía en un estadio

[46]JONDEC BRIONES, H. P. et al., "La necesidad de regulación del deepfake en el ordenamiento jurídico", Vniversitas Jurídica, vol. 73 (2024), pp. 1-28.

incipiente. Las empresas tecnológicas y los proveedores de sistemas de IA generativa deben asumir obligaciones legales de asistencia a las autoridades judiciales, facilitando la trazabilidad de los flujos de datos que alimentan las infraestructuras delictivas. Modelos de colaboración como la plataforma EMPACT de Europol[49] o los mecanismos de intercepción de pagos I-GRIP de INTERPOL —que han permitido recuperar cientos de millones de dólares en operaciones recientes—[50] marcan la dirección estratégica que debe orientar la acción institucional para el período 2025-2027.

Adicionalmente, resulta necesario dotar a las unidades especializadas de los cuerpos de seguridad —como el Centro Europeo de Ciberdelincuencia (EC3) de Europol— de un mandato operativo que les permita desplegar capacidades de IA contrarrestante (*counter-AI*) para infiltrarse en los ecosistemas criminales y desmantelar las plataformas de *Deepfake-as-a-Service* en tiempo real, bajo control judicial y con plenas garantías de los derechos fundamentales.[51] El objetivo final para esta década debe ser la construcción de una infraestructura de confianza digital en la que la autenticidad de las interacciones sea verificable tanto técnica como jurídicamente, garantizando que el progreso tecnológico no opere en detrimento de los derechos fundamentales reconocidos en la CDFUE y en el Convenio Marco del Consejo de Europa.[52]

[49]EUROPOL, EMPACT (European Multidisciplinary Platform Against Criminal Threats), Ciclo de política 2022-2025, prioridad OCP7 sobre fraude financiero digital, La Haya, 2022.
[50]INTERPOL, I-GRIP (Incident Response and Global Recovery Platform), resultados operativos 2024, comunicado de prensa, 15 de enero de 2025.
[51]EL PACCTO 2.0, op. cit., pp. 55-60; EUROPOL EC3, "Counter-AI capabilities in cybercrime investigation", documento de trabajo interno citado en EL PACCTO 2.0.

BIBLIOGRAFÍA

I. Normativa

Carta de los Derechos Fundamentales de la Unión Europea, DO C 326, 26 de octubre de 2012.

Convenio Marco sobre Inteligencia Artificial y Derechos Humanos, Democracia y Estado de Derecho, CETS núm. 225, Consejo de Europa, abierto a la firma el 5 de septiembre de 2024.

Convención de Budapest sobre la Ciberdelincuencia, CETS núm. 185, Consejo de Europa, Budapest, 23 de noviembre de 2001.

Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence (EO 14110), 30 de octubre de 2023, Federal Register, vol. 88, núm. 210.

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

ISO/IEC 30107-3:2023, Information technology — Biometric presentation attack detection — Part 3: Testing and reporting, International Organization for Standardization, Ginebra, 2023.

Proyecto de Ley Orgánica de protección de las personas menores de edad en los entornos digitales, BOCG, Congreso de los Diputados, Serie A, núm. 51-1, XIV Legislatura (2023).

Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial [AI Act / RIA], DO L, 12 de julio de 2024.

Regulations on the Management of Deep Synthesis Technology, Administración del Ciberespacio de China, en vigor desde el 10 de enero de 2023.

II. Jurisprudencia

STJUE de 13 de mayo de 2014, Google Spain SL y Google Inc. c. AEPD y Mario Costeja González, C-131/12, EU:C:2014:317.

STS (Sala de lo Penal) de 21 de octubre de 2021, recurso núm. 1167/2020 (documento informático como documento a efectos del art. 26 CP).

III. Doctrina

ACALE SÁNCHEZ, María & BOZA MARTÍNEZ, Diego, "La imagen ante el Derecho penal: nuevos desafíos", e-Eguzkilore, Revista del Instituto Vasco de Criminología, núm. 18 (2025), pp. 3-22.

BLÁZQUEZ MORENO, R., "Deepfakes en el procedimiento probatorio", Instituto Vasco de Derecho Procesal, 2023.

ESTELLA, A., "Regulación de las 'ultrasuplantaciones' en el Reglamento de la UE sobre IA", Revista de Administración Pública, núm. 221 (2023), pp. 89-120.

GOODFELLOW, Ian et al., "Generative Adversarial Networks", Advances in Neural Information Processing Systems, vol. 27 (2014), pp. 2672-2680.

JONDEC BRIONES, H. P. et al., "La necesidad de regulación del deepfake en el ordenamiento jurídico", Vniversitas Jurídica, vol. 73 (2024), pp. 1-28.

NASH, John F., "Equilibrium Points in N-Person Games", Proceedings of the National Academy of Sciences, vol. 36, núm. 1 (1950), pp. 48-49.

RIBAS, Xavier, "Ataques de presentación mediante deepfakes en videoconferencia", Ribas Artículos, 2024.

SÁNCHEZ SALAZAR, P. M. et al., "Responsabilidad penal por manipulación digital con fines delictivos", Revista Cuestiones Políticas, vol. 40, núm. 73 (2022), pp. 212-235.

WANG, Yuxuan et al., "Neural Voice Cloning with a Few Samples", Advances in Neural Information Processing Systems, vol. 31 (2018).

IV. Documentos institucionales e informes

AI Incident Database, Incident 634 / Report 3642: Arup deepfake fraud, disponible en: https://incidentdatabase.ai.

CEIM (Confederación Empresarial de Madrid-CEOE), "Informe sobre ciberseguridad y su impacto en las empresas", 2025.

Cybersecurity Law Report, "Deepfake fraud in global firms: Ferrari, WPP and Arup case studies", julio de 2025.

Deloitte Center for Financial Services, "The coming wave of AI-enabled financial fraud", 2024.

EL PACCTO 2.0 (Expertise France / FIAP), "Weaponizing Artificial Intelligence: How AI Reshapes the World of Organized Crime", 2025.

EUROPOL, "ChatGPT — The impact of Large Language Models on Law Enforcement", EC3 SPOTLIGHT Report, La Haya, 2023.

EUROPOL, EMPACT (European Multidisciplinary Platform Against Criminal Threats), Ciclo de política 2022-2025, La Haya, 2022.

Group-IB Threat Intelligence, "Hi-Tech Crime Trends 2024", Singapur, 2024.

INTERPOL, "Financial Crimes Threat Assessment 2024", INTERPOL General Secretariat, Lyon, 2024.

INTERPOL, I-GRIP (Incident Response and Global Recovery Platform), resultados operativos 2024, comunicado de prensa, 15 de enero de 2025.

NIST, AI Risk Management Framework (AI RMF 1.0), National Institute of Standards and Technology, Gaithersburg, 2023.

GLOSARIO DE TÉRMINOS TÉCNICO-JURÍDICOS

Algoritmos opacos
Modelos de inteligencia artificial cuyo proceso interno de toma de decisiones no resulta auditable por medios técnicos convencionales. La opacidad de estos modelos dificulta la trazabilidad del engaño y, con ello, la atribución de responsabilidad penal en el marco de los artículos 27 y 28 del Código Penal español.

Ataque de presentación
Tentativa maliciosa de engañar a un sistema de reconocimiento biométrico o a un interlocutor humano mediante la exhibición de una identidad manipulada. Regulado técnicamente por la norma ISO/IEC 30107-3:2023.

Deepfake / Ultrasuplantación
Contenido de imagen, audio o vídeo generado o manipulado mediante técnicas de inteligencia artificial que induce a error sobre la veracidad de lo que una persona dice o hace. A los efectos del Reglamento (UE) 2024/1689, constituye un contenido sintético sujeto a las obligaciones de transparencia del artículo 50.

Face-swapping (intercambio de rostros)
Técnica que emplea redes neuronales para superponer el rostro de una persona sobre el de otra en un flujo de vídeo, logrando un grado de realismo suficiente para la suplantación en videoconferencias en tiempo real.

Lege ferenda (de)
Locución latina que designa el derecho tal como debería ser creado o reformado, por contraposición a la lex lata o derecho vigente. Empleada en este informe para encuadrar las propuestas de reforma legislativa.

Redes Adversarias Generativas (GANs)

Arquitectura de inteligencia artificial compuesta por dos redes neuronales —generadora y discriminadora— que compiten entre sí para producir contenidos sintéticos tendencialmente indistinguibles de los contenidos auténticos. Conceptualizadas por Goodfellow et al. en 2014.

SGSI (Sistema de Gestión de Seguridad de la Información)
Marco organizativo y técnico basado en el estándar ISO/IEC 27001:2022 que permite a las organizaciones gestionar sistemáticamente los riesgos asociados a la seguridad de la información.

Vibe hacking
Forma avanzada de ingeniería social en la que la inteligencia artificial imita no solo el aspecto físico o la voz de un individuo, sino su estilo comunicativo, sus patrones de expresión y su urgencia emocional característica, con el fin de anular el juicio crítico de la víctima.

Voice cloning (clonación de voz)
Técnica de síntesis de voz que replica el timbre, la entonación y el acento de una persona a partir de muestras mínimas de audio, conforme al estado de la técnica descrito por Wang et al. (2018).

Zero Trust (confianza cero)
Filosofía de ciberseguridad y principio de arquitectura de sistemas que parte de la premisa de que ninguna interacción —interna o externa— puede considerarse auténtica por defecto, exigiendo la autenticación continua de todos los usuarios, dispositivos y transacciones.