Fraude en plataformas de contenidos digitales: análisis multidisciplinar, vectores de ataque y modelo de mitigación

Dr. Juan José Sánchez Peña

Doctor en Informática. Director del Máster de Ciberseguridad UAX

Dr. Gonzalo Martínez Ginesta

Doctor en Seguridad de la Información.

Dr. Alejandro Corletti Estrada

Doctor en Informática.

Dr. José Antonio Marcos García

Doctor en Informática. Vicedecano de Tecnología UAX

Grupo de Investigación en Ciberseguridad y Tecnologías Aplicadas

Facultad de Business & Tech

Universidad Alfonso X el Sabio (UAX)

Abstract (ES)

El crecimiento exponencial de la economía digital y la consolidación de las plataformas de contenido como infraestructuras clave han generado un entorno propicio para la evolución del fraude digital. Este fenómeno ha transitado desde manifestaciones aisladas hacia un sistema complejo, escalable y altamente sofisticado, que combina vulnerabilidades técnicas, mecanismos financieros y estrategias de manipulación cognitiva.

Este trabajo analiza el fraude digital en plataformas de contenido desde una perspectiva multidisciplinar, abordando su conceptualización, tipologías y vectores de ataque. Se examinan categorías como el robo de credenciales, la suplantación de identidad, los esquemas financieros fraudulentos y la manipulación de información, así como el papel de tecnologías habilitadoras emergentes como la inteligencia artificial. Asimismo, se estudia el ecosistema de plataformas digitales y sus modelos de negocio como configuradores de superficies de ataque específicas.

A partir de este análisis, se identifican implicaciones estratégicas, destacando la industrialización del fraude, la convergencia intersectorial y la erosión de la confianza digital. El trabajo propone un modelo de actuación basado en prevención, detección, intervención y persecución, junto con un conjunto de recomendaciones orientadas a fortalecer la resiliencia del ecosistema digital.

Metodológicamente, el estudio se fundamenta en la revisión de literatura académica y sectorial, así como en el análisis de informes de organismos internacionales y evidencia empírica reciente. El objetivo es proporcionar un marco analítico y operativo que contribuya al diseño de estrategias coordinadas de mitigación del fraude digital en España. Para ello, el trabajo propone el Modelo Socio-Técnico del Fraude en Plataformas de Contenido (MSFPC), un marco conceptual que formaliza el fraude como un sistema adaptativo multi-capa y desarrolla un conjunto de hipótesis contrastables sobre las relaciones entre complejidad de la superficie de ataque, sofisticación del fraude, capacidad adaptativa de los atacantes e impacto en el ecosistema digital. Complementariamente, se articula un modelo operativo de actuación basado en cuatro ejes —prevención, detección, intervención y persecución (PDIP)— junto con diez recomendaciones estratégicas orientadas al contexto institucional español.

Palabras clave—Fraude digital; plataformas de contenido; cibercrimen; ingeniería social; inteligencia artificial; deepfakes; seguridad digital.

Abstract (EN)

The exponential growth of the digital economy and the consolidation of content platforms as key infrastructures have created an environment conducive to the evolution of digital fraud. This phenomenon has transitioned from isolated incidents to a complex, scalable, and highly sophisticated system that combines technical vulnerabilities, financial mechanisms, and cognitive manipulation strategies.

This paper analyzes digital fraud in content platforms from a multidisciplinary perspective, addressing its conceptualization, typologies, and attack vectors. It examines categories such as credential theft, identity impersonation, financial fraud schemes, and information manipulation, as well as the role of enabling technologies such as automation and artificial intelligence. Additionally, the digital platform ecosystem and its business models are analyzed as key drivers of specific attack surfaces.

Based on this analysis, the paper identifies strategic implications for the Spanish context, highlighting the industrialization of fraud, cross-sector convergence, and the erosion of digital trust. It proposes an operational framework based on prevention, detection, intervention, and enforcement, along with a set of recommendations aimed at strengthening the resilience of the digital ecosystem.

Methodologically, the study is based on a review of academic and industry literature, as well as the analysis of international reports and recent empirical evidence. The objective is to provide both an analytical and operational framework to support the development of coordinated strategies to mitigate digital fraud in Spain. To this end, the paper proposes the Socio-Technical Model of Fraud in Content Platforms (MSFPC), a conceptual framework that formalises fraud as a multi-layered adaptive system and develops a set of testable hypotheses on the relationships between attack surface complexity, fraud sophistication, attacker adaptive capacity, and ecosystem impact. Additionally, it articulates an operational response model structured around four axes —prevention, detection, intervention, and enforcement (PDIP)— together with ten strategic recommendations tailored to the Spanish institutional context.

Keywords—Digital fraud; content platforms; cybercrime; social engineering; artificial intelligence; deepfakes; digital security.

Índice de contenidos

Índice de figuras

Índice de tablas

0. Preámbulo

A pesar del creciente volumen de literatura académica sobre cibercrimen y fraude digital, la investigación existente presenta diversas limitaciones a la hora de abordar el fraude en plataformas de contenido. En primer lugar, la mayoría de los estudios se centran bien en la dimensión técnica (p.ej.: malware o ataques a redes), bien en el fraude financiero (p.ej.: sistemas de pago), sin ofrecer una perspectiva integrada que conecte los factores tecnológicos, conductuales y económicos. En segundo lugar, existe una articulación limitada del fraude como fenómeno sistémico e intersectorial, especialmente en lo que respecta a la convergencia entre telecomunicaciones, servicios financieros y plataformas digitales. En tercer lugar, la literatura actual carece con frecuencia de marcos operativos que traduzcan los hallazgos analíticos en estrategias accionables para las instituciones públicas.

Esta laguna resulta especialmente relevante en el contexto de las estrategias nacionales de ciberseguridad, donde las instituciones no solo requieren análisis descriptivos, sino también modelos estructurados para la intervención y la coordinación. En consecuencia, existe una necesidad real de investigación que tienda un puente entre el análisis académico y la aplicación orientada a la política pública.

Contribución del estudio

Este estudio contribuye a la literatura existente en tres dimensiones principales: (i) un marco analítico integrado que combina perspectivas técnicas, económicas y conductuales; (ii) un enfoque sistémico que conceptualiza el fraude como un ecosistema industrializado e intersectorial; y (iii) un modelo operativo orientado a la política pública, articulado en cuatro ejes de actuación y diez recomendaciones estratégicas. Estas contribuciones se desarrollan en detalle en las Secciones 11.1 y 11.2.

Metodología

Este estudio adopta un diseño de investigación cualitativo y analítico, orientado a la construcción de un marco conceptual integrador a partir de la síntesis de fuentes heterogéneas. La elección de este enfoque responde a la naturaleza del objeto de estudio: el fraude digital en plataformas de contenido es un fenómeno multidimensional, dinámico y distribuido entre múltiples actores y jurisdicciones, cuya comprensión requiere articular dimensiones técnicas, económicas, conductuales y regulatorias que la literatura existente ha abordado de forma fragmentada. Un enfoque cuantitativo basado en datos primarios resultaría prematuro sin un marco teórico que defina previamente los constructos relevantes y sus relaciones — precisamente lo que este trabajo propone.

Estrategia de búsqueda y selección de fuentes

La investigación se fundamenta en cuatro categorías de fuentes, seleccionadas en función de su relevancia, actualidad y trazabilidad:

Primera, literatura académica: publicaciones indexadas en bases de datos de referencia (Scopus, Web of Science, IEEE Xplore, SSRN) sobre cibercrimen, fraude digital, ingeniería social, sistemas socio-técnicos y seguridad de la información. Los criterios de búsqueda combinaron términos como *digital fraud*, *content platform*, *social engineering*, *fraud detection*, *socio-technical systems* y *adaptive security*, con filtros de relevancia y fecha de publicación (preferentemente 2018-2026, con inclusión de obras seminales anteriores como Anderson (2020), Cialdini (2009), Bostrom y Heinen (1977) o Baxter y Sommerville (2011) por su carácter fundacional).

Segunda, informes institucionales y sectoriales: documentos publicados por organismos de referencia en el ámbito de las telecomunicaciones (GSMA Fraud and Security Group, CFCA), la ciberseguridad (ENISA, INCIBE, CCN-CERT, NCSC), la regulación financiera (Banco de España, BCE), la cooperación policial (Europol) y organizaciones internacionales (UNODC, World Economic Forum, Global Initiative Against Transnational Organized Crime). Se priorizaron los informes correspondientes al periodo 2023-2026, con atención especial a las sesiones FASG#33 (2025) y FASG#34 (2026) de la GSMA, cuyos datos empíricos sobre vectores de ataque, impacto económico y patrones operativos constituyen una de las bases factuales principales del trabajo.

Tercera, legislación y normativa: textos regulatorios de la Unión Europea (RGPD, DSA, NIS2, EU AI Act, PSD2/PSD3), legislación española (Código Penal, Esquema Nacional de Seguridad) y estándares sectoriales relevantes para la contextualización del marco legal del fraude digital.

Cuarta, fuentes digitales especializadas: publicaciones de medios sectoriales, blogs tecnológicos corporativos y análisis de empresas de ciberseguridad (Kaspersky, McAfee, BioCatch, ESET), seleccionadas cuando aportan datos empíricos verificables, estadísticas de incidencia o análisis técnicos de vectores de ataque no disponibles en la literatura académica convencional. La inclusión de estas fuentes responde a la velocidad de evolución del fenómeno estudiado, donde la evidencia más reciente se publica frecuentemente en canales sectoriales antes de su formalización académica.

Marco temporal

El núcleo de la revisión cubre el periodo 2023-2026, coincidiendo con la aceleración del fraude asociado a inteligencia artificial generativa, la entrada en vigor de la DSA (2024) y la proximidad de la fecha de aplicación del EU AI Act (agosto 2026). Las fuentes anteriores a 2023 se incluyen cuando aportan marcos teóricos fundacionales o datos de referencia para la contextualización longitudinal del fenómeno.

Enfoque analítico

El análisis se estructura en tres niveles complementarios. En el primer nivel, descriptivo-taxonómico, se identifican y clasifican las tipologías de fraude, los vectores de ataque y los mecanismos de defensa existentes (Capítulos 2 a 6). En el segundo nivel, analítico-comparativo, se examinan las implicaciones estratégicas del fenómeno para el contexto español y se formulan recomendaciones operativas (Capítulos 4.9, 7). En el tercer nivel, teórico-propositivo, se desarrolla el Modelo Socio-Técnico del Fraude en Plataformas de Contenido (MSFPC), que formaliza las relaciones entre constructos mediante un modelo

causal con trece hipótesis contrastables, y se articula el modelo operativo PDIP (Capítulo 9).

Construcción de casos

El Capítulo 8 presenta tres casos end-to-end construidos mediante la técnica de caso tipificado (Yin, 2018), en la que se sintetizan patrones operativos recurrentes documentados en múltiples fuentes primarias (GSMA FASG#33/34, Europol, Ministerio del Interior, Kaspersky, McAfee) para construir narrativas representativas del fenómeno. Los casos no corresponden a incidentes individuales identificables, sino a tipologías compuestas que reflejan las cadenas de ataque más frecuentes en el ecosistema europeo y español. Esta aproximación permite ilustrar la dinámica del fraude sin comprometer información operativa sensible ni incurrir en atribuciones no verificables.

1. Introducción y marco conceptual

Este capítulo establece el marco conceptual del estudio. Se estructura en dos secciones: una delimitación conceptual del fraude cibernético, el cibercrimen y los delitos informáticos, y una contextualización del fraude específico en plataformas de contenido digital, incluyendo su impacto económico, sobre los usuarios y social.

1.1. Introducción al fraude digital: definición de fraude cibernético, cibercrimen y delitos informáticos

El término **fraude cibernético** alude a aquellas conductas ilícitas que se desarrollan mediante el uso de sistemas informáticos, redes y plataformas digitales con el fin de obtener un beneficio económico, reputacional o de control indebido sobre personas, empresas o infraestructuras (Europol, 2025). En el contexto de las plataformas de contenido digital, el fraude cibernético incluye, entre otros, el robo de credenciales, la suplantación de identidad, el uso fraudulento de suscripciones, la manipulación de sistemas de pago y la obtención de ingresos ilícitos a través de publicidad falsa o tráfico engañoso (Global Initiative Against Transnational Organized Crime, 2026; Kaspersky, 2025). Estas prácticas se distinguen del uso legítimo de la tecnología por su intencionalidad engañosa, su violación de normas legales y su capacidad para generar daño material o reputacional a las víctimas.

El **cibercrimen**, en cambio, constituye un concepto más amplio que engloba cualquier cometido ilícito por medios digitales, abarcando tanto delitos de carácter económico como de naturaleza social, política o de seguridad nacional (Europol, 2025; UNODC, 2013). En el ámbito de plataformas de contenido, el cibercrimen se manifiesta en la distribución de malware, el uso de redes sociales para la captación de víctimas, el alojamiento de contenido ilegal (pornografía infantil, incitación al odio, terrorismo) y la explotación de servicios de streaming y marketplaces como infraestructura de ataque (Kaspersky, 2025; Global Initiative Against Transnational Organized Crime, 2026). La especificidad del cibercrimen reside en su dependencia estructural de la red y de las plataformas digitales, lo que multiplica la escala, la velocidad y la dificultad de rastreo de los actos delictivos (UNODC, 2013).

Por su parte, los **delitos informáticos** son aquellas conductas típicas reprobadas por el Derecho penal que tienen como soporte o vector primario un sistema informático, una red de datos o un dispositivo digital (Gobierno de España, 2023; UNODC, 2013). En la legislación española, por ejemplo, el Código Penal tipifica como delitos informáticos la intrusión indebida en sistemas informáticos, la interrupción de sistemas mediante ataques de denegación de servicio, la suplantación de identidad digital y la manipulación de datos protegidos, siempre que se realicen con ánimo de lucro, daño o peligro para la seguridad de terceros (Gobierno de España, 2023). En el contexto de plataformas de contenido, estos delitos se actualizan cuando se accede a cuentas de usuario sin autorización, se modifica la configuración de suscripciones, se alteran los registros de consumo o se roban bases de datos de usuarios para su comercialización en mercados negros (Global Initiative Against Transnational Organized Crime, 2026; Digital Innovation News, 2026).

Si bien existe una superposición entre fraude cibernético, cibercrimen y delitos informáticos, resulta útil distinguir:

- **Fraude cibernético**: se centra en el engaño como mecanismo de obtención de beneficio (por ejemplo, suplantación de Netflix, plataformas de inversión falsas, SpyLoan).
- **Cibercrimen**: abarca todos los usos ilícitos de la infraestructura digital, incluidos los no monetarios (acoso, difusión de contenido ilegal, ciberterrorismo).
- **Delitos informáticos**: se refieren a la categoría jurídica específica dentro del ordenamiento penal, muchas veces aplicable a las dos categorías anteriores cuando se tipifican conductas como acceso indebido, suplantación o alteración de datos (Gobierno de España, 2023).

En el entorno de plataformas de contenido digital, estas tres nociones convergen en un mismo ecosistema de superficie de ataque: cuentas de usuario, mecanismos de autenticación, sistemas de pago, bases de datos de clientes y algoritmos de recomendación. Por ello, muchos casos detectados se encuadran simultáneamente como fraude cibernético (desde la perspectiva económica), cibercrimen (desde la óptica de la seguridad pública) y delito informático (desde la lente penal) (Global Initiative Against Transnational Organized Crime, 2026; Europol, 2025). Esta superposición refuerza la necesidad de abordajes interdisciplinares que combinen el análisis técnico de la seguridad, la regulación de plataformas y el diseño de políticas de prevención y respuesta tanto a nivel nacional como internacional (UNODC, 2013).

1.2. El fraude en plataformas de contenido digital: contextualización e impacto económico y social

La economía digital ha transformado fundamentalmente la forma en que la sociedad accede consume y comparte contenidos, reconfigurando la relación entre ciudadanos, empresas y Estado en un entorno altamente interconectado (Banco de España, 2025; Hispania Segura, 2025). Plataformas de streaming de vídeo y música, redes sociales, servicios de contenido bajo demanda y marketplaces de aplicaciones se han consolidado como espacios centrales de la vida cotidiana, tanto en el ámbito del ocio como en el profesional y educativo (Banco de España, 2025; World Economic Forum, 2025). Sin embargo, este crecimiento exponencial de la digitalización ha generado una correlación directa con el aumento de actividades delictivas que explotan vulnerabilidades técnicas, humanas y organizativas en los sistemas digitales, lo que configura un escenario de fraude digital sistémico y globalizado (Global Initiative Against Transnational Organized Crime, 2026; Visa España, 2025).

En España, el fenómeno del fraude digital se ha consolidado como una de las principales amenazas a la estabilidad económica y social contemporánea. Datos oficiales correspondientes al primer trimestre de 2025 revelan que se han registrado 106.800 infracciones penales vinculadas a estafas informáticas, lo que representa un incremento cercano al 40 % respecto al mismo periodo del año anterior, según reportes de las Fuerzas y Cuerpos de Seguridad del Estado (Europa Press, 2025; InfoBae, 2025b). Esta cifra equivale aproximadamente a 1.200 estafas online por día en el territorio español, lo que evidencia la escala de la actividad delictiva y la dificultad para su contención mediante mecanismos tradicionales (Atresmedia, 2025; Europa Press, 2025).

El impacto económico directo del fraude digital en España es especialmente significativo. Un estudio elaborado por Visa en 2025, ampliamente citado por el Banco de España,

señala que el fraude digital genera pérdidas superiores a 350 millones de euros anuales para la economía española, afectando a consumidores, empresas y servicios financieros (Visa España, 2025; PressDigital, 2025). Diversos medios y análisis sectoriales recogen que, en 2025, el coste medio anual del fraude digital supera ya los 350 millones de euros, mientras que algunas estimaciones alternativas sitúan el impacto global en torno a 500 millones de euros anuales, reflejando la magnitud de la exposición del sistema financiero y de pago nacional (Atresmedia, 2025; InfoBae, 2025b). A nivel social, el auge del fraude deteriora la confianza en los mercados digitales, reduce la voluntad de adoptar servicios online (particularmente en el comercio electrónico y en la banca digital) y fomenta la desinformación sobre la seguridad de las tecnologías (Digital Innovation News, 2026; Management Society, 2025).

A escala global, el impacto económico del fraude digital y de los fenómenos asociados a la desinformación se sitúa en niveles alarmantes. Estudios internacionales recientes señalan que la desinformación generada y amplificada por inteligencia artificial, junto con el fraude digital en plataformas de contenido, impone un coste estimado de 78.000 millones de dólares anuales a la economía mundial (World Economic Forum, 2025; Revista Mercado, 2025). Esta cifra no se limita a pérdidas monetarias directas en forma de transferencias fraudulentas o pagos no autorizados, sino que incluye también daños indirectos como la volatilidad bursátil inducida por noticias falsas, la erosión de confianza en instituciones, el deterioro de la efectividad de políticas públicas y la reducción de la inversión empresarial en sectores afectados por ciberdelincuencia sistémica (World Economic Forum, 2025; El Debate, 2024).

Un informe global de 2026 sobre “Un mundo de engaños” subraya que el fraude digital supera ya el billón de dólares anuales en ingresos ilícitos, consolidándose como una de las actividades más lucrativas del crimen organizado a escala internacional (Global Initiative Against Transnational Organized Crime, 2026). Este volumen de negocio ilícito se nutre precisamente de la fragmentación del mercado digital, la disponibilidad de herramientas de automatización y la dificultad de coordinación regulatoria entre jurisdicciones (Global Initiative Against Transnational Organized Crime, 2026; McAfee, 2025). En este contexto, plataformas de contenido digital —como servicios de streaming, redes sociales y marketplaces— funcionan a la vez como canales de difusión de contenidos y como infraestructuras de ataque para phishing, suplantación de identidad, distribución de software malicioso y captación de víctimas para estafas financieras y modelos delictivos como el “SpyLoan” (malware oculto en aplicaciones falsas de préstamos rápidos para Android) o el blanqueo de capitales mediante cuentas mulas (Kaspersky, 2025; Ministerio del Interior, 2026).

La fragmentación del mercado de plataformas de contenido digital ha intensificado estas tendencias delictivas. En 2024, los sitios de piratería registraron más de 216.000 millones de visitas, un incremento notable respecto a los 130.000 millones de 2020, lo que evidencia la reactivación de modelos de consumo ilegal impulsados por la saturación de ofertas legales, la subida de precios y la brecha de accesibilidad geográfica o económica (Hispania Segura, 2025; Spain Audiovisual Hub, 2025). Esta dinámica no solo supone una pérdida de ingresos para las industrias creativas y de contenidos, sino que también expone a millones de usuarios a riesgos de ciberataques, malware, suplantación de identidad y robo de credenciales, ya que muchas plataformas pirata carecen de mecanismos de seguridad adecuados y funcionan como superficie de ataque para redes criminales organizadas (Hispania Segura, 2025; Sénal News, 2025).

El impacto en el sector cultural y creativo es especialmente relevante. La piratería digital y el fraude de suscripciones afectan a productoras audiovisuales, sellos discográficos, estudios de videojuegos y plataformas de ocio en línea, generando un desplazamiento de valor desde la creación legítima hacia el mercado negro (Spain Audiovisual Hub, 2025; Coalición de Creadores e Industrias de Contenidos, 2025). En 2025, organizaciones sectoriales como la Coalición de Creadores e Industrias de Contenidos señalan que la piratería se mantiene en niveles históricos, lo que reduce la capacidad de inversión en nuevos contenidos y la remuneración adecuada de autores e intérpretes (Coalición de Creadores e Industrias de Contenidos, 2025). A nivel macroeconómico, esto se traduce en una menor capacidad de exportación de marcas y contenidos españoles, sumando al impacto directo del fraude digital en el consumo interno (Visa España, 2025; Digital Innovation News, 2026).

A nivel social, el fraude digital en plataformas de contenido contribuye a la polarización informativa, la desconfianza hacia las instituciones y la fragmentación del discurso público. La difusión de fake news, deepfakes y contenido manipulado en redes sociales no solo afecta a la reputación de personas y empresas, sino que también interfiere con procesos democráticos, campañas electorales y la calidad del debate público (World Economic Forum, 2025; El Debate, 2024). Usuarios vulnerables, como personas mayores o grupos con menor alfabetización digital, se ven especialmente expuestos a estafas de phishing, suplantación de identidad y engaños financieros, lo que agrava desigualdades sociales y genera un efecto de exclusión digital (Digital Innovation News, 2026; Management Society, 2025). Informes recientes sobre desinformación corporativa subrayan que, en 2025, más de la mitad de las grandes empresas españolas ha sufrido directamente los efectos de noticias falsas vinculadas a su marca, reforzando la dimensión reputacional del fraude digital (El Debate, 2024).

En conjunto, el fraude digital en plataformas de contenido se configura como un problema multimodal: combate la innovación económica, deteriora la confianza social en la tecnología, tensiona los sistemas legales y regulatorios y redistribuye de forma ilícita el valor generado por la economía digital (Global Initiative Against Transnational Organized Crime, 2026; World Economic Forum, 2025). Tanto en España como a nivel global, la urgencia de respuestas coordinadas entre sector público, sector privado y sociedad civil se vuelve evidente, ya que la digitalización de la vida cotidiana continúa acelerando la superficie de ataque y la sofisticación de las estrategias criminales (Banco de España, 2025; McAfee, 2025). En 2025-2026, la combinación de modelo de servicios de suscripción, crecimiento de las finanzas digitales y evolución de la inteligencia artificial aplicada a la manipulación de información define un escenario en el que las plataformas de contenido pasan a ser, a la vez, eje de la economía digital y vector central de la ciberdelincuencia (World Economic Forum, 2025; Global Initiative Against Transnational Organized Crime, 2026).

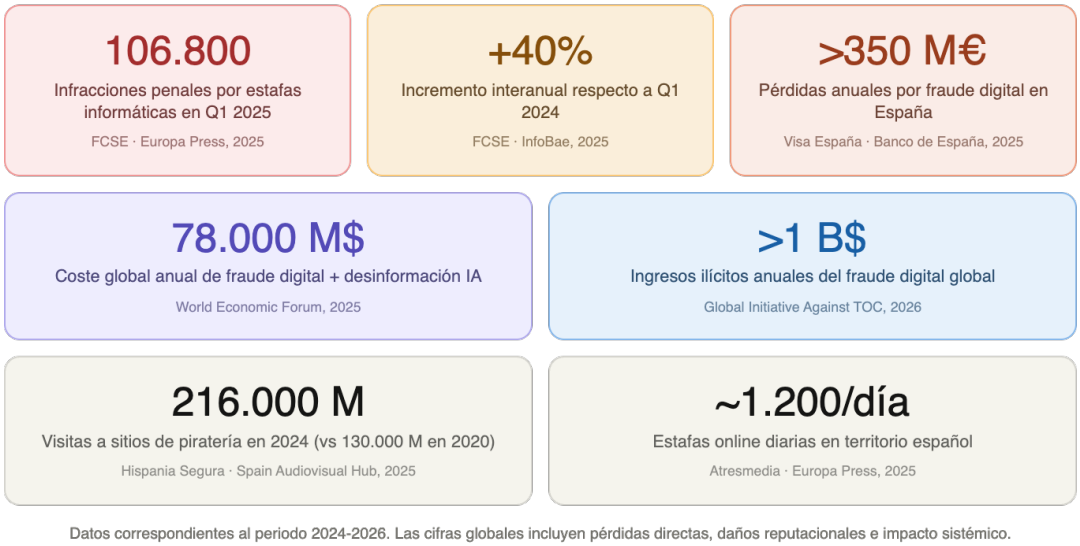

Figura 1 Dimensión del fraude digital: principales indicadores de impacto en España y a escala global (2024-2026). Elaboración propia a partir de FCSE (Europa Press, 2025), Visa España (2025), WEF (2025), Global Initiative Against Transnational Organized

A continuación se profundiza en las dimensiones específicas del impacto del fraude sobre el ecosistema de plataformas de contenido.

1.2.1. Impacto económico

El fraude produce pérdidas directas derivadas del uso no autorizado de servicios y del acceso ilícito a contenidos. Este impacto se intensifica en un contexto de transformación estructural del sector audiovisual, caracterizado por la disminución sostenida de modelos tradicionales de TV de pago y el crecimiento acelerado del vídeo online. Según datos recientes del sector, las suscripciones de TV de pago en Norteamérica muestran un declive continuado mientras que las suscripciones de vídeo online crecen de forma sostenida, con previsiones que sitúan la base de suscriptores de streaming significativamente por encima de la de pay-TV para 2030 (Omdia, 2025).

La evolución hacia modelos híbridos de monetización, que combinan suscripción con publicidad (SVOD/AVOD), incrementa la exposición al fraude al multiplicar los puntos de interacción económica dentro del ecosistema digital. En 2025, los ingresos publicitarios agregados de los servicios híbridos SVOD/AVOD en Estados Unidos superan por primera vez a los del mercado combinado de AVOD/FAST/BVOD, lo que refleja un desplazamiento estructural del valor hacia modelos de streaming con publicidad (Omdia, 2025). Este contexto multiplica las oportunidades de fraude publicitario, fraude de suscripción y acceso no autorizado a contenidos.

En mercados maduros, la saturación de infraestructuras de fibra y la consiguiente migración de hogares desde paquetes de TV tradicional hacia planes de banda ancha con servicios OTT a la carta está acelerando la recomposición del sector. En este entorno, la industria media atraviesa un período crítico de transformación que incluye escisiones corporativas, fusiones de plataformas y optimización de carteras de contenidos (Omdia, 2025), lo que genera tanto oportunidades como vulnerabilidades para actores fraudulentos.

A nivel global, estimaciones recientes sitúan el impacto del fraude en mensajería y plataformas digitales en decenas de miles de millones de dólares anuales, con casos específicos que alcanzan decenas de millones en una sola operación (GSMA FASG#34, 2026). Este impacto debe entenderse en el contexto de la escalabilidad descrita en la Sección 4.2, donde la capacidad de ejecución masiva amplifica el alcance de los ataques.

En el ámbito de mobile money, se estima que un operador medio puede perder aproximadamente 1,06 millones de dólares anuales debido al fraude, lo que impacta directamente en su rentabilidad y competitividad (GSMA FASG#33, 2025). A nivel sectorial, las pérdidas por fraude en telecomunicaciones representan aproximadamente un 5% de los ingresos anuales, un porcentaje que incluye fraude de bypass (SIMBox), IRSF, roaming fraud, Wangiri y PABX hacking, entre otros esquemas (GSMA FASG#33, 2025).

Además, existen costes indirectos asociados a mitigación, litigios, recuperación de sistemas y deterioro reputacional (Anderson, 2020). Los marcos operativos del sector reconocen la necesidad de cuantificar estos impactos mediante indicadores clave de rendimiento (KPIs) que combinen métricas aplicadas a servicios de voz, SMS y datos con indicadores específicos de fraude en TV, contenidos y piratería. Como se ha indicado en la Sección 4.7, la industrialización del fraude amplifica estos efectos a escala sistémica.

1.2.2. Impacto sobre los usuarios

Los usuarios son simultáneamente víctimas y vectores del fraude digital. Los principales impactos incluyen:

- **Robo de identidad**, frecuentemente ejecutado mediante técnicas de SIM swapping que permiten al atacante interceptar comunicaciones y acceder a cuentas bancarias, redes sociales y plataformas de contenido.
- **Filtración de datos personales**, facilitada por la explotación de vulnerabilidades en protocolos de señalización (SS7) o por campañas masivas de phishing dirigidas a usuarios de plataformas de streaming.
- **Acceso indebido a cuentas**, que puede resultar en el secuestro de perfiles de usuario, modificación de datos de pago y utilización no autorizada de servicios contratados.
- **Suscripción involuntaria a servicios premium no deseados**, derivada de manipulaciones técnicas en la facturación directa al operador (Direct Carrier Billing).
- **Robo de credenciales**, **fraude financiero y distribución de malware** derivados de ataques mediante SMS blasters y estaciones base falsas, que operan fuera del control del operador y explotan canales percibidos como fiables (GSMA FASG#34, 2026).

Más allá del daño económico, el fraude genera efectos psicológicos como pérdida de confianza y percepción de inseguridad (Cross, 2018). La investigación en este ámbito indica que las víctimas de fraude digital experimentan niveles de estrés y ansiedad comparables a los de otros delitos contra la persona, particularmente cuando el ataque afecta a la identidad digital o a la privacidad de las comunicaciones.

A nivel conductual, el fraude genera una pérdida de confianza en los canales digitales, especialmente en el SMS, que históricamente ha sido percibido como un canal fiable. En entornos de mobile money, estos efectos se traducen en incrementos del churn de hasta un 33% y reducciones del volumen transaccional de hasta un 66%, lo que convierte al fraude en un factor estratégico que condiciona la sostenibilidad del modelo de negocio (GSMA FASG#33, 2025).

Esta erosión de la confianza afecta directamente a la adopción de servicios digitales, estableciendo un vínculo directo con las dinámicas de ingeniería social descritas en la Sección 4.1.

1.2.3. Impacto social y desplazamiento del fraude

A nivel macro, el fraude digital contribuye a la degradación del ecosistema informativo. La proliferación de plataformas digitales incrementa la complejidad del entorno: según datos de la industria, en determinados mercados europeos coexisten más de 230 servicios de streaming, lo que dificulta a los usuarios la identificación de servicios legítimos frente a plataformas fraudulentas o de contenido pirateado (Omdia, 2025).

La consolidación del sector mediante fusiones y adquisiciones —ilustrada por la escisión de grandes conglomerados mediáticos en compañías separadas de streaming y redes lineales, o por la integración de plataformas domésticas en mercados competitivos como el surcoreano— reconfigura constantemente el paisaje de servicios disponibles, generando períodos de confusión que los actores fraudulentos explotan activamente.

El fraude presenta además una capacidad significativa de adaptación frente a medidas regulatorias. Evidencia empírica muestra que la introducción de mecanismos como registros de Sender ID puede provocar el desplazamiento del fraude hacia otros canales, como P2P o short codes (ComReg, 2025; GSMA FASG#34, 2026). Este fenómeno sugiere que las intervenciones parciales pueden tener efectos no deseados, desplazando el problema en lugar de resolverlo. En términos más amplios, la proliferación de fraude contribuye a la erosión de la confianza en el ecosistema digital, afectando tanto a usuarios como a proveedores de servicios.

Cuando se combina con tecnologías como la inteligencia artificial, el fraude facilita la desinformación y la manipulación de la opinión pública (Wardle & Derakhshan, 2017). Las mismas técnicas de deepfake empleadas para el fraude financiero pueden utilizarse para generar contenido audiovisual falso atribuido a medios de comunicación o figuras públicas, erosionando la confianza en la información online y la estabilidad de los entornos digitales.

Este fenómeno, vinculado a las capacidades descritas en la Sección 4.3, plantea riesgos significativos que trascienden el ámbito económico y afectan al funcionamiento democrático de las sociedades.

2. Ecosistema de plataformas de contenido digital

Este capítulo analiza el ecosistema de plataformas de contenido digital, examinando la tipología de plataformas existentes y los modelos de negocio que configuran su superficie de ataque.

El ecosistema de plataformas de contenido digital constituye un entorno complejo y en constante evolución, en el que convergen servicios, modelos de negocio y dinámicas de interacción muy diversas. Su relevancia no se limita al consumo de contenidos, sino que se extiende a ámbitos económicos, sociales y tecnológicos cada vez más interdependientes. En este contexto, resulta necesario analizar tanto la tipología de plataformas como su arquitectura funcional y sus principales superficies de ataque, con el fin de comprender cómo se articulan en ellas las oportunidades para la comisión de fraude digital.

Este capítulo examina, por tanto, las distintas clases de plataformas que integran dicho ecosistema, así como los modelos de negocio que las sustentan y los vectores de riesgo asociados a cuentas de usuario, suscripciones, sistemas de pago, publicidad y tratamiento de datos personales. A partir de esta aproximación, se pretende ofrecer una visión estructurada que permita vincular la expansión de la economía de plataformas con las nuevas formas de ciberdelincuencia que la acompañan.

2.1. Tipología de plataformas: streaming de vídeo/música, redes sociales, servicios de contenido bajo demanda, marketplaces de apps y contenidos

El ecosistema de plataformas de contenido digital es hoy un conjunto heterogéneo y altamente fragmentado, en el que conviven servicios de entretenimiento, comunicación, comercio y creación de contenidos bajo una misma lógica de acceso mediante internet (InnovaOrgen, 2025; Xataka, 2026).

En este contexto, se distinguen, al menos, cinco grandes familias de plataformas: servicios de streaming de vídeo y música, redes sociales, plataformas de contenido bajo demanda (VOD/OTT), marketplaces de aplicaciones móviles y marketplaces de contenidos digitales (audiovisual, software, ebooks, etc.) (ResearchNester, 2026; Xataka, 2026).

Cada una de estas tipologías presenta su propia arquitectura tecnológica, modelo de negocio, superficie de ataque y lógica de interacción con el usuario, lo que configura un escenario de fraude digital diferenciado pero interconectado (Kaspersky, 2025; Global Initiative Against Transnational Organized Crime, 2026).

Los servicios de streaming de vídeo y música se caracterizan por ofrecer acceso on-demand a contenidos audiovisuales o musicales mediante suscripción, pago por uso o publicidad integrada (Xataka, 2026; ResearchNester, 2026). En el ámbito del vídeo, plataformas como Netflix, Disney+, HBO Max, Prime Video, Movistar+ y plataformas de nicho como Filmin o Apple TV+ dominan el mercado occidental, combinando catálogos extensos de series, películas, documentales y, en algunos casos, canales de televisión en directo (Xataka, 2026). En el ámbito de la música, servicios como Spotify, Apple Music, Amazon Music, Deezer y plataformas de nicho como Tidal configuran un

mercado global que se valora en más de 56.000 millones de dólares en 2025 y proyectado para superar los 200.000 millones en 2035 (ResearchNester, 2026; Stereojoint, 2024). Estos servicios se basan en abonos recurrentes, perfiles de usuario, recomendaciones personalizadas mediante inteligencia artificial y políticas de acceso simultáneo limitadas, elementos que se convierten a la vez en vectores de fraude de suscripción, compartición ilícita de cuentas, robo de credenciales y suplantación de perfil (Kaspersky, 2025; Xataka, 2026).

Las redes sociales constituyen un segundo gran bloque de plataformas de contenido digital, centradas en la interacción, la conversación y la difusión de contenido audiovisual y multimedia (Agencia Comma, 2025; Limón Publicidad, 2025). En 2025-2026, plataformas como Facebook, Instagram, TikTok, X (Twitter), LinkedIn, YouTube y Twitch dominan el panorama comunicativo, con diferentes perfiles de uso: Facebook e Instagram para interacción social y marketing de marca, TikTok para video-short bajo demanda y viralidad, LinkedIn para networking profesional y Twitch y YouTube Live para streaming en vivo de gaming y contenidos de entretenimiento (Agencia Comma, 2025; Limón Publicidad, 2025; Xataka, 2026). En este entorno proliferan modelos como el social commerce (Instagram Shop, TikTok Shop), que integran pago y recomendación de productos dentro de la propia plataforma, ampliando la superficie de ataque para estafas de phishing, anuncios engañosos, perfiles falsos y distribución de malware o deepfakes orientados a la captación de víctimas (Kaspersky, 2025; Limón Publicidad, 2025).

Un tercer grupo lo conforman los servicios de contenido bajo demanda (VOD/OTT), que incluyen tanto plataformas de streaming especializadas como servicios de vídeo y música bajo demanda que se distribuyen de forma “over-the-top”, es decir, sin necesidad de pasar por operadores de telecomunicaciones tradicionales (FractalMedia, 2024). A partir de 2025, el mercado OTT se proyecta alcanzar cerca de 1.99 billones de dólares en 2029, impulsado por la fragmentación de catálogos, el auge de contenidos exclusivos y la competencia entre conglomerados mediáticos (FractalMedia, 2024). En este marco se incluyen plataformas verticales especializadas en fitness, cocina, educación, bienestar, fanbases deportivos o religiosos, que combinan modelos de suscripción, alquiler de contenido e incluso publicidad segmentada para mantener ingresos sin depender únicamente de la publicidad convencional (FractalMedia, 2024; Stereojoint, 2024). La arquitectura de estas plataformas, con microcuentas, flujos de pago heterogéneos y sistemas de autenticación cruzada, se convierte en un terreno fértil para el fraude de pago, la manipulación de métricas de consumo y la usurpación de identidad para el acceso no autorizado a contenidos restringidos (Kaspersky, 2025; Global Initiative Against Transnational Organized Crime, 2026).

Un cuarto grupo está formado por los marketplaces de aplicaciones móviles, es decir, tiendas de apps como Google Play, Apple App Store, Huawei AppGallery y tiendas alternativas, donde se distribuyen aplicaciones de entretenimiento, productividad, redes sociales, juegos y servicios financieros (Kaspersky, 2025). Estos ecosistemas permiten a desarrolladores publicar apps de pago, freemium o con compras internas, generando una superficie de ataque para apps falsas, troyanos que se disfrazan de servicios legítimos, compras in-app fraudulentas y estafas de suscripción no deseadas (Kaspersky, 2025; Global Initiative Against Transnational Organized Crime, 2026). El crecimiento de plataformas de streaming en vivo integradas en apps móviles (Twitch, Instagram Live, TikTok Live, YouTube Live) añade una capa adicional de riesgo, ya que combinan microtransacciones, donaciones, regalos virtuales y canales de comunicación directos

entre creadores y audiencia, que pueden usarse para captar víctimas en estafas o esquemas de blanqueo de capitales mediante cuentas mulas (Kaspersky, 2025; Sénal News, 2025).

Finalmente, los marketplaces de contenidos digitales agrupan plataformas dedicadas a la distribución y venta de software, juegos, ebooks, música, cursos online y otros productos digitales, tanto en formato de suscripción como de compra única (AceleraPyme, 2024; Future Market Insights, 2025). En el ámbito de videojuegos, plataformas como Steam, Epic Games Store, GOG y Xbox Game Pass funcionan como marketplaces de contenidos digitales, gestionando transacciones, suscripciones, regalos de licencias y almacenamiento en la nube de progresos de usuario (Kaspersky, 2025). De forma análoga, plataformas de cursos online (Coursera, Udemy, Domestika), bibliotecas digitales y servicios de música de nicho permiten la adquisición de contenidos mediante tarjeta de crédito, billeteras digitales o suscripciones recurrentes, lo que abre la puerta a fraudes de cuenta, compras no autorizadas, clonación de tarjetas y modelos de estafa tipo “SpyLoan” que exploten vulnerabilidades en los flujos de pago y autenticación (Forbes Business Council, 2024; Global Initiative Against Transnational Organized Crime, 2026).

En conjunto, la tipología de plataformas de contenido digital se configura como un ecosistema multi-plataforma en el que cada grupo (streaming, redes sociales, OTT, tiendas de apps y marketplaces de contenidos) comparte patrones de fraude digitales comunes —como el robo de credenciales, la suplantación de identidad y la manipulación de sistemas de pago—, aunque con lógicas específicas derivadas de su modelo de negocio y arquitectura (Kaspersky, 2025; Global Initiative Against Transnational Organized Crime, 2026). Esta diversidad tipológica impone un diseño de contramedidas diferenciado, que combine políticas de autenticación robusta, monitorización de transacciones, educación de usuarios y regulación transversal, ya que el uso simultáneo de múltiples plataformas por parte de usuarios individuales multiplica la exposición a riesgos de fraude digital (InnovaOrgen, 2025; Xataka, 2026).

2.2. Modelos de negocio y superficie de ataque: cuentas de usuario, suscripciones, sistemas de pago, publicidad y datos personales

Los modelos de negocio de las plataformas de contenido digital se configuran, en 2024-2026, en torno a una combinación de suscripciones recurrentes, publicidad, microtransacciones y ventas de contenidos, que generan flujos de ingresos altamente dependientes del volumen de usuarios activos y de la calidad de los datos personales que gestionan (Kaspersky, 2025; Global Initiative Against Transnational Organized Crime, 2026).

En streaming de vídeo y música, el modelo clásico es el abono mensual por cuenta, con opciones de familia o grupos de usuarios, políticas de acceso simultáneo limitadas y sistemas de recomendación basados en datos de comportamiento (Xataka, 2026; ResearchNester, 2026).

En redes sociales, el modelo se centra en la interacción, la viralidad y la monetización de anuncios, influencer marketing y funcionalidades de e-commerce social (Agencia Comma, 2025; Limón Publicidad, 2025).

En marketplaces de apps y contenidos digitales domina una mezcla de venta de licencias, suscripciones y compras internas, que se gestionan a través de tiendas móviles y plataformas de pago integradas (Forbes Business Council, 2024; Kaspersky, 2025).

Estos modelos de negocio definen, al mismo tiempo, la superficie de ataque de los ciberdelincuentes, cuya estrategia se orienta a explotar cada uno de los puntos de contacto entre el usuario, la plataforma y el sistema de pago (Global Initiative Against Transnational Organized Crime, 2026; UNODC, 2013).

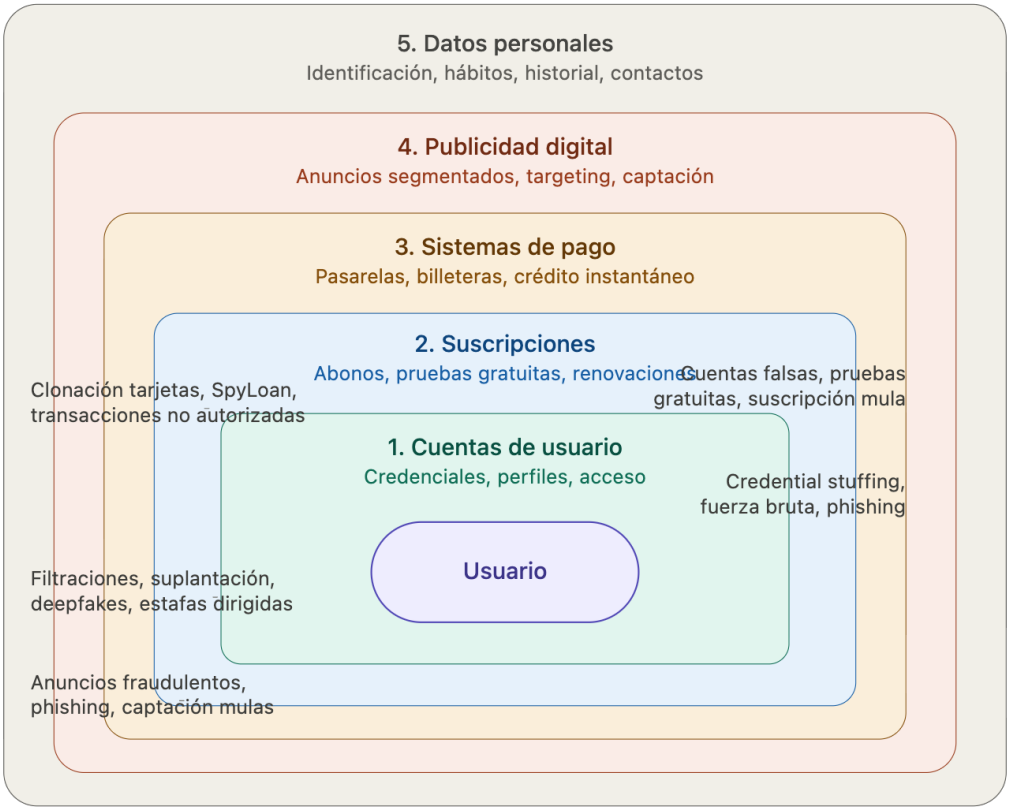

Figura 2 Las cinco capas de superficie de ataque en plataformas de contenido digital: cuentas de usuario, suscripciones, sistemas de pago, publicidad y datos personales, con vectores de explotación asociados. Elaboración propia.

La cuenta de usuario constituye el primer nivel de superficie de ataque: a través del robo de credenciales, phishing, ataques de fuerza bruta o credential stuffing, los ciberdelincuentes obtienen acceso a perfiles de streaming, redes sociales y plataformas de contenido, desde los que pueden consumir contenidos, comprar productos, interactuar con otros usuarios o incluso alojar material ilegal (Kaspersky, 2025; Digital Innovation News, 2026). La compartición fraudulenta de cuentas y la venta de credenciales robadas en foros y mercados negros amplían este riesgo, al permitir a múltiples actores hacer uso simultáneo de una misma cuenta, dificultando la detección y aumentando la exposición de datos personales y de pago (Global Initiative Against Transnational Organized Crime, 2026; Sénal News, 2025).

El modelo de suscripción representa una segunda superficie de ataque clave. Muchas plataformas gestionan abonos automáticos vinculados a tarjetas de crédito, billeteras digitales o registros de pago recurrente, lo que convierte cada alta, cambio o baja de

suscripción en un vector potencial de fraude (Visa España, 2025; Europol, 2025). Los ciberdelincuentes manipulan sistemas de registro para crear cuentas masivas con identidades falsas, registrar tarjetas de crédito robadas o explotar promociones de prueba gratuita, generando pérdidas de ingresos para las plataformas y exponiendo a los titulares reales de la tarjeta a riesgos de rescate de datos y carga de responsabilidad económica (Kaspersky, 2025; Global Initiative Against Transnational Organized Crime, 2026). Además, la suplantación de promociones, la creación de campañas de suscripción falsas y la utilización de cuentas "mulas" para captar suscripciones fraudulentas se han vuelto frecuentes en marketplaces de apps y servicios de contenido digital (Ministerio del Interior, 2026; Digital Innovation News, 2026).

Los sistemas de pago y la infraestructura financiera asociada constituyen una tercera superficie de ataque esencial. Plataformas de contenido integran pasarelas de pago, billeteras, suscripciones recurrentes y, en algunos casos, modelos de compra impulsiva vinculados a redes sociales o apps móviles (Forbes Business Council, 2024; Kaspersky, 2025). Este diseño favorece la experiencia de usuario, pero también multiplica las posibilidades de compras no autorizadas, clonación de tarjetas, transacciones de prueba y esquemas de SpyLoan que explotan datos de pago almacenados o el acceso a cuentas de usuario (Global Initiative Against Transnational Organized Crime, 2026; UNODC, 2013). La integración de sistemas de crédito instantáneo, microcréditos o "préstamos" en apps falsas, que se difunden a través de redes sociales y marketplaces de apps, refuerza la lógica de fraude financiero basado en datos personales y tarjetas de crédito comprometidas (Kaspersky, 2025; Sénal News, 2025).

La publicidad digital funcionando en redes sociales, marketplaces y plataformas de streaming constituye una cuarta superficie de ataque estratégica. Muchas plataformas generan ingresos mediante la venta de anuncios segmentados, cuya eficacia depende de la calidad, cantidad y precisión de los datos de usuarios, lo que incentiva la recolección masiva de información personal, comportamental y de localización (Limón Publicidad, 2025; Agencia Comma, 2025). Sin embargo, esta misma infraestructura permite la difusión de anuncios fraudulentos, enlaces maliciosos, campañas de phishing y mensajes de estafa financiera, que se dirigen a usuarios con perfil de consumo alto o baja alfabetización digital (Kaspersky, 2025; Digital Innovation News, 2026). El uso de técnicas de targeting refinado permite a los ciberdelincuentes automatizar la captación de víctimas para modelos de fraude clásico y nuevos formatos como el blanqueo de capitales mediante cuentas "mulas" (Ministerio del Interior, 2026; UNODC, 2013).

En paralelo, la gestión de datos personales constituye una quinta superficie de ataque crítica. Las plataformas de contenido recopilan, procesan y almacenan datos de identificación, hábitos de consumo, historial de búsqueda, patrones de pago y contactos sociales de los usuarios, lo que aumenta enormemente el valor de un perfil de usuario comprometido (Global Initiative Against Transnational Organized Crime, 2026; Europol, 2025). El robo o la filtración de bases de datos de usuarios permite a los ciberdelincuentes llevar a cabo suplantación de identidad, estafas personalizadas, deepfakes y ataques de ingeniería social dirigidos a empleadores, bancos o familiares, intensificando el impacto del fraude digital más allá de la propia plataforma (Kaspersky, 2025; UNODC, 2013). La dificultad de aplicar controles de protección de datos homogéneos en un ecosistema multi-plataforma, donde cada servicio tiene su propia política de privacidad y su propio modelo de gestión de consentimiento, agrava la exposición de los usuarios al fraude y a la manipulación de información (InnovaOrgen, 2025; Xataka, 2026).

En conjunto, los modelos de negocio de plataformas de contenido digital configuran una superficie de ataque segmentada en cinco grandes capas: cuentas de usuario, suscripciones, sistemas de pago, publicidad y gestión de datos personales (Kaspersky, 2025; Global Initiative Against Transnational Organized Crime, 2026). Cada una de estas capas ofrece a los ciberdelincuentes un conjunto de vectores concretos de explotación, que se articulan en torno a la obtención de beneficios económicos, el acceso a información sensible o el control sobre la identidad digital de las víctimas (UNODC, 2013; Europol, 2025). Frente a este escenario, la respuesta efectiva requiere un diseño de seguridad y gobernanza que integre la protección de datos, la autenticación robusta, la monitorización de transacciones y los sistemas de pago, al tiempo que se reconoce que el ecosistema multi-plataforma hace que el fraude digital en una sola superficie se propague con rapidez al resto del tejido digital (InnovaOrgen, 2025; Kaspersky, 2025).

3. Tipología de fraudes en plataformas de contenido y marco conceptual

El presente capítulo analiza la tipología de fraudes que afectan a las plataformas de contenido digital, atendiendo a sus principales modalidades, mecanismos de ejecución y lógica de funcionamiento. A lo largo de este apartado se observa cómo estas conductas delictivas no responden a un único patrón, sino que se despliegan sobre distintas capas del ecosistema digital, desde el acceso a las cuentas y la manipulación de identidades hasta la explotación de marcas, sistemas de pago y contenidos informativos. Esta diversidad obliga a abordar el fraude digital como un fenómeno estructural, transversal y adaptado a las características de cada plataforma.

En este marco, se examinan cuatro grandes grupos de fraude: los fraudes de acceso y cuenta, los fraudes de contenido y marca, los fraudes financieros asociados y los fraudes basados en la manipulación de información. Esta clasificación permite ordenar el análisis y poner de relieve tanto los elementos comunes que comparten estas prácticas como sus particularidades operativas, jurídicas y preventivas. De este modo, el capítulo ofrece una visión integrada de las principales amenazas que afectan al ecosistema de plataformas de contenido digital y de su progresiva sofisticación en el entorno actual.

Figura 3 Taxonomía del fraude en plataformas de contenido digital: cuatro categorías principales y vectores específicos, con técnicas transversales del Capítulo 4. Elaboración propia.

3.1. Fraudes de acceso y cuenta: robo de credenciales, apropiación de cuentas, compartición fraudulenta de cuentas, venta de cuentas robadas

Los fraudes de acceso y cuenta constituyen una de las manifestaciones más frecuentes y dañinas del fraude digital en plataformas de contenido, ya que se dirigen directamente al núcleo de la relación entre usuario y servicio: la identidad digital y el acceso autorizado (Kaspersky, 2025; Global Initiative Against Transnational Organized Crime, 2026). Este tipo de fraude se centra en el deterioro de los mecanismos de autenticación, la gestión de perfiles de usuario y los sistemas de autorización, lo que permite a los ciberdelincuentes obtener beneficios sin el consentimiento de la víctima, bien a través de consumo indebido de contenidos, robo de datos personales o utilización de la cuenta como superficie de ataque para terceros (Europol, 2025; UNODC, 2013).

El robo de credenciales es, probablemente, la modalidad más extendida dentro de esta categoría. A través de técnicas como phishing, keylogging, credential stuffing y ataques de fuerza bruta, los ciberdelincuentes obtienen contraseñas, identificadores de usuario y, en muchos casos, datos de pago vinculados a cuentas de streaming, redes sociales, marketplaces y plataformas de contenido (Kaspersky, 2025; Digital Innovation News, 2026). Phishing masivo, que se difunde mediante correos electrónicos, mensajes SMS, anuncios en redes sociales o páginas web falsas que imitan servicios legítimos, induce a los usuarios a introducir sus credenciales en formularios de acceso fraudulentos, que se almacenan y se revenden en mercados negros (Kaspersky, 2025; Global Initiative Against Transnational Organized Crime, 2026). Por su parte, el credential stuffing aprovecha listas de credenciales obtenidas en filtraciones previas para probarlas de forma automatizada en múltiples plataformas, aprovechando que muchos usuarios reutilizan contraseñas en diferentes servicios (Europol, 2025).

La apropiación de cuentas, una vez obtenidas las credenciales, supone la toma de control efectivo del perfil de usuario, lo que permite a los ciberdelincuentes modificar datos personales, cambiar métodos de pago, desactivar la autenticación de dos factores y ocultar rastros de su actividad (Digital Innovation News, 2026; Kaspersky, 2025). En plataformas de streaming, redes sociales y servicios de contenido bajo demanda, este tipo de fraude se traduce en consumo indebido de contenidos, publicación de contenido fraudulento o malicioso, utilización de la cuenta como punto de difusión de estafas o malware, o incluso suplantación de la identidad de la víctima ante contactos, familiares o empleadores (Global Initiative Against Transnational Organized Crime, 2026; UNODC, 2013). La dificultad de recuperar la cuenta, unida a la lentitud de respuesta de algunos servicios, agrava el impacto emocional y reputacional sobre la víctima, además del daño económico derivado de cargos no autorizados (Digital Innovation News, 2026).

Otro tipo de fraude de acceso se centra en **la compartición fraudulenta de cuentas y la venta de cuentas robadas** en mercados y foros online (Global Initiative Against Transnational Organized Crime, 2026; Sénal News, 2025).

Muchos usuarios comparten de forma voluntaria credenciales de streaming, redes sociales o servicios de contenido con familiares o amigos, en el marco de modelos de "cuenta compartida" diseñados por las propias plataformas; sin embargo, la misma práctica se extiende de forma ilícita cuando se comercializan cuentas robadas, perfiles premium o suscripciones a precios reducidos en chats de mensajería, foros de piratería o

marketplaces de acceso digital (Sénal News, 2025; Xataka, 2026). En estos esquemas, los ciberdelincuentes generan ingresos multiplicando el número de usuarios que acceden a una misma cuenta, muchas veces pagando la suscripción inicial con tarjetas de crédito robadas o datos falsos, lo que genera riesgos de rescate de cuenta, bloqueos masivos y tensiones con los proveedores de pago (Kaspersky, 2025; Visa España, 2025).

La venta de cuentas robadas se integra en un ecosistema distinto de cibermercados online, donde se negocian perfiles de usuario completos, incluyendo historial de consumo, datos de pago, contactos y, en ocasiones, niveles de acceso privilegiados (Global Initiative Against Transnational Organized Crime, 2026; UNODC, 2013). Estos perfiles, una vez adquiridos por otros ciberdelincuentes, se utilizan para la realización de uno o varios de los siguientes objetivos: consumo de contenido de pago sin autorización, captación de nuevas víctimas a través de la lista de contactos o conversaciones previas, robo de datos de pago adicionales o utilización de la cuenta como plataforma de crimen organizado (por ejemplo, blanqueo de capitales, redes de mulas digitales o distribución de software malicioso) (Kaspersky, 2025; Ministerio del Interior, 2026). La lógica de mercado que regula estos esquemas —con precios variables según el nivel de acceso, la antigüedad de la cuenta o la presencia de métodos de pago asociados— refuerza la dimensión económica del fraude de acceso y cuenta, convirtiendo la identidad digital en un bien a comercializar en el mercado negro (UNODC, 2013).

En el contexto de plataformas de contenido, la combinación de robo de credenciales, apropiación de cuentas, compartición fraudulenta y venta de cuentas robadas genera un efecto sistémico de inseguridad: reduce la percepción de confianza en los servicios digitales, tensiona los sistemas de autenticación y obliga a los proveedores a implementar medidas más intrusivas (como verificaciones de identidad más estrictas o sistemas de pago recurrente más complejos), lo que, a su vez, puede afectar la experiencia de usuario (Kaspersky, 2025; Digital Innovation News, 2026). Al mismo tiempo, estas prácticas favorecen la normalización de comportamientos de riesgo entre los usuarios, que a menudo perciben la compartición de cuentas o el uso de credenciales “regaladas” como un mecanismo de ahorro sin comprender plenamente las implicaciones legales y de seguridad asociadas (Global Initiative Against Transnational Organized Crime, 2026; Europol, 2025).

En conjunto, los fraudes de acceso y cuenta se configuran como un conjunto de prácticas orientadas a la explotación de la superficie de autenticación y autorización de las plataformas digitales, convirtiendo la identidad y el acceso en el vector principal de la ciberdelincuencia en el ecosistema de contenidos digitales (Global Initiative Against Transnational Organized Crime, 2026; Kaspersky, 2025). La respuesta a este tipo de fraude exige una combinación de tecnologías de autenticación avanzada (como autenticación multifactor y sistemas biométricos), monitorización de accesos anómalos, educación de usuarios y regulación que penalice la comercialización de cuentas robadas y la reutilización de credenciales de terceros, para reducir la superficie de ataque y restaurar la confianza en la economía digital (Europol, 2025; UNODC, 2013).

3.2. Fraudes de contenido y marca: suplantación de identidad, perfiles falsos, anuncios engañosos, phishing y enlaces maliciosos distribuidos vía redes sociales

Los fraudes de contenido y marca se sitúan en el cruce entre la reputación de empresas y marcas digitales y la percepción de autenticidad de los usuarios, generando un escenario en el que el contenido mismo se convierte en arma o vehículo de delincuencia (EN THEC, 2025; Cybersecurity News, 2026). A diferencia de los fraudes centrados únicamente en cuentas y sistemas de pago, este tipo de estafas explota directamente la confianza en la marca, el reconocimiento visual y la expectativa de legitimidad que los usuarios asocian a servicios conocidos como Netflix, Spotify, Facebook, WhatsApp o Amazon, entre otros (Kaspersky, 2025; ESET, 2025). En este contexto, la suplantación de identidad, la creación de perfiles falsos, los anuncios engañosos y la difusión de enlaces maliciosos mediante redes sociales se configuran como las modalidades más relevantes de fraude de contenido y marca en plataformas digitales (EN THEC, 2025; InfoBae, 2025a).

La suplantación de identidad es una de las estrategias centrales de este tipo de fraude. Los ciberdelincuentes generan páginas, perfiles o mensajes que imitan a servicios legítimos, utilizando logos, tipografías, colores corporativos y estructuras de lenguaje deliberadamente parecidas a las de las marcas originales (Kaspersky, 2025; EN THEC, 2025). Estas imitaciones se dirigen a la confianza cognitiva del usuario, que tiende a reconocer rápidamente símbolos visuales familiares y aceptar sin cuestionar el origen del contenido (MiTek Systems, 2025). El objetivo suele ser el robo de credenciales, datos de pago, información biométrica o la ejecución de acciones que beneficien a los atacantes, como la descarga de malware, la autorización de accesos remotos o la participación en campañas de "verificación KYC (Know Your Customer – Conoce a tu cliente" falsa (Kaspersky, 2026; ESET, 2025).

Un segundo vector de fraude de marca lo constituyen **los perfiles falsos** en redes sociales, que copian nombres, fotos y biografías de personas reales o de empresas para establecer comunicación con usuarios y captar sus datos o dinero (InfoBae, 2025a; Cybersecurity News, 2026). Estos perfiles se construyen a partir de información pública de los usuarios, se mantienen activos mediante interacción simulada y, en muchos casos, incorporan verificaciones automatizadas o artificiales para reforzar la sensación de legitimidad (Kaspersky, 2025; SecureList, 2025). Una vez que se establece contacto con la víctima, los perfiles falsos pueden iniciar conversaciones de amistad, buscar favores económicos, ofrecer trabajos falsos o promover productos y servicios que no existen más allá de la simulación digital (EN THEC, 2025; Cybersecurity News, 2026). En algunos casos, se utilizan para replicar la imagen de ejecutivos, celebridades o figuras públicas con el fin de influir en decisiones financieras o de reputación (Kaspersky, 2025; Red Seguridad, 2025).

Otro tipo de fraude de contenido se centra en **los anuncios engañosos y** campañas de **publicidad falsa** que se difunden a través de redes sociales, buscadores y plataformas de contenido (EN THEC, 2025; Cybersecurity News, 2026). Estas campañas, que se clasifican como malvertising, consisten en anuncios que aparentan promocionar productos legítimos, ofertas exclusivas, pruebas de software o servicios de suscripción, pero que en realidad redirigen a páginas fraudulentas o descargas maliciosas (Kaspersky, 2025; Cybersecurity News, 2026; SecureList, 2025). En 2025, se observó un incremento significativo de estafas relacionadas con compras online, tiendas falsas y promociones de

"última oportunidad" difundidas mediante anuncios pagados en Facebook, YouTube y otras plataformas, aprovechando la confianza que los usuarios depositan en los canales de publicidad (EN THEC, 2025; Cybersecurity News, 2026). La lógica de segmentación de dichas plataformas permite dirigir mensajes de alto impacto emocional (por ejemplo, descuentos extraordinarios o inventarios agotados) a usuarios con perfil de consumo elevado, lo que multiplica la eficacia de las estafas (EN THEC, 2025; InfoBae, 2025a).

Un tercer grupo de fraude de contenido y marca se centra en el **phishing y enlaces maliciosos** distribuidos vía redes sociales, que se integran en la dinámica comunicativa de plataformas como Facebook, Instagram, X (antigua Twitter), WhatsApp, Telegram o TikTok (Kaspersky, 2025; ESET, 2025; SecureList, 2025). El phishing en redes sociales puede adoptar múltiples formas: mensajes directos que simulan correspondencia de servicios conocidos, enlaces cortos que ocultan la URL real, anuncios que promueven supuestas verificaciones de seguridad o invitaciones para participar en sorteos o encuestas que requieren el ingreso de credenciales (Kaspersky, 2025; ESET, 2025; SecureList, 2025). Estos ataques se han vuelto cada vez más personalizados, **gracias al uso de inteligencia artificial** para analizar perfil y comportamiento de los usuarios, lo que permite a los atacantes diseñar mensajes altamente creíbles y adaptados al contexto de cada víctima (SecureList, 2025; Cybersecurity News, 2026). El resultado es que muchos usuarios no perciben diferencia visible entre un mensaje legítimo y otro fraudulentamente diseñado, lo que facilita la obtención de datos sensibles o la autorización involuntaria de accesos a servicios de terceros (Kaspersky, 2025; ESET, 2025).

En el entorno de plataformas de contenido, **la suplantación de marcas** se extiende también a la creación de tiendas falsas y servicios de streaming pirata que se presentan como plataformas oficiales o alternativas económicas, pero que en realidad operan sin autorización y sin garantías de protección de datos (EN THEC, 2025; Cybersecurity News, 2026). Estas tiendas, muchas de las cuales se promocionan a través de anuncios pagados o campañas de influencer marketing, se aprovechan de la saturación de ofertas legítimas y de la búsqueda de soluciones más económicas por parte de los usuarios para generar ingresos fraudulentos y captar suscripciones que no se traducen en verdaderos servicios de contenido (EN THEC, 2025; Cybersecurity News, 2026). El daño se extiende no solo a los consumidores, que pierden dinero y exponen sus datos, sino también a las marcas legítimas, cuya reputación se ve afectada por la vinculación con contenidos y servicios falsos (EN THEC, 2025; MiTek Systems, 2025).

Finalmente, el fraude de contenido y marca se ha intensificado con la incorporación de **deepfakes** y generación sintética de voz y vídeo, que permiten a ciberdelincuentes suplantar a ejecutivos, figuras públicas o empleados de empresas para realizar estafas financieras, extorsión o manipulación de información (Red Seguridad, 2025; Kaspersky, 2025; Cybersecurity News, 2026). En 2024 tuvo lugar una de las primeras grandes estafas deepfake, en la que se utilizó una videollamada falsa para simular la presencia de un ejecutivo y autorizar una transferencia de millones de dólares (Red Seguridad, 2025). Desde entonces, se ha documentado un aumento significativo de esquemas en los que la imagen o la voz de una persona real se replica con alta precisión para engañar a otros empleados, a clientes o a terceros proveedores, lo que agrava el riesgo de compromiso de la identidad digital a nivel corporativo y personal (Kaspersky, 2025; Cybersecurity News, 2026; UI1, 2025).

En conjunto, los fraudes de contenido y marca se configuran como un conjunto de prácticas orientadas a la explotación de la percepción de legitimidad que los usuarios

asocian a plataformas, marcas y figuras públicas (EN THEC, 2025; Cybersecurity News, 2026). A través de la suplantación de identidad, la creación de perfiles falsos, la difusión de anuncios engañosos y la manipulación de phishing y enlaces maliciosos, estos fraudes buscan desplazar la confianza del usuario de la marca legítima hacia entidades digitales diseñadas exclusivamente para su aprovechamiento (Kaspersky, 2025; ESET, 2025). La respuesta a este tipo de fraude exige una combinación de medidas técnicas (como monitoreo de perfiles, detección de deepfakes y filtros de phishing), estrategias de comunicación de marca y concienciación del usuario para que pueda reconocer y denunciar señales de suplantación, tanto en plataformas de contenido como en redes sociales y servicios de pago (EN THEC, 2025; Cybersecurity News, 2026).

3.3. Fraudes financieros asociados: apps falsas, plataformas de inversión fraudulentas, modelos “SpyLoan”, mulas digitales y blanqueo.

Los fraudes financieros asociados a plataformas de contenido digital se configuran como un conjunto de esquemas diseñados para obtener beneficios económicos directos sobre el patrimonio de los usuarios, combinando tecnologías móviles, infraestructuras de pago y la lógica de redes sociales y marketplaces (Kaspersky, 2025; Cybersecurity News, 2025).

En este contexto, destaca la proliferación de aplicaciones falsas, plataformas de inversión fraudulentas, modelos de crédito maliciosos tipo “SpyLoan”, el uso de cuentas mulas digitales y la articulación de estos elementos en dinámicas de blanqueo de capitales, lo que refuerza la interconexión entre ciberdelincuencia, finanzas y reputación digital (ESET, 2023; BioCatch, 2024).

Las apps falsas representan una de las vías más extendidas de fraude financiero. A través de tiendas de apps oficiales y alternativas, los ciberdelincuentes distribuyen aplicaciones que simulan servicios legítimos, como bancos, billeteras, herramientas de pago, alarmas de seguridad o servicios de streaming, pero cuya verdadera finalidad es el robo de credenciales, datos de pago o información sensible del dispositivo (Kaspersky, 2024; McAfee, 2024). Muchas de estas apps se integran en recomendaciones de tiendas preinstaladas, anuncios patrocinados o campañas en redes sociales, lo que confiere una apariencia de legitimidad que favorece la instalación masiva (Kaspersky, 2024; ESET, 2023). Una vez instaladas, estas aplicaciones pueden ejecutar funciones de spyware para monitorizar ingresos de contraseñas, tomar capturas de pantalla, leer mensajes o robar contactos, lo que genera un daño financiero y reputacional sobre el usuario y sirve de punto de partida para otros tipos de fraude, como el phishing o el blanqueo de capitales (Kaspersky, 2024; McAfee, 2024).

Un segundo grupo de fraude financiero se centra en **las plataformas de inversión fraudulentas y las aplicaciones de inversión falsas**, que prometen rendimientos altos, sin riesgo y en plazos breves, explotando la creciente atención hacia criptomonedas, divisas, materias primas y bienes inmuebles (Sin Embargo, 2025; Kaspersky, 2024). Estas plataformas se presentan con diseño profesional, testimonios falsos, gráficos manipulados y apariencia de respaldo institucional, lo que induce a los usuarios a invertir ahorros en activos inexistentes o ficticios (Sin Embargo, 2025; Kaspersky, 2024). La Policía Cibernética de la Ciudad de México alertó en 2025 sobre el auge de apps y plataformas falsas de inversión, muchas de las cuales se promocionan a través de redes sociales,

mensajería instantánea y correos electrónicos, utilizando nombres similares a instituciones financieras legítimas y figuras públicas para reforzar la credibilidad (Sin Embargo, 2025; Kaspersky, 2024). En muchos casos, los usuarios no recuperan su dinero una vez realizado el depósito, al tiempo que exponen información personal y bancaria susceptible de ser utilizada para otros delitos (Sin Embargo, 2025).

Un tercer tipo de fraude financiero lo representan **los modelos "SpyLoan", aplicaciones de préstamos maliciosas** que se distribuyen en plataformas móviles y redes sociales bajo la promesa de créditos rápidos, flexibles y con mínimos requisitos (ESET, 2023; McAfee, 2023). Aunque aparentan ofrecer préstamos con condiciones favorables, en la práctica estas apps se diseñan para obtener el máximo de información posible del usuario, incluyendo contactos, fotos, historial de navegación y datos de pago (Kaspersky, 2024; McAfee, 2023). Cuando el usuario no cumple con los pagos o se incurre en incumplimientos deliberadamente generados por el propio sistema, se inicia una fase de acoso y extorsión, que puede incluir bloqueo del dispositivo, envío de mensajes a contactos para presionar con el pago o amenazas de divulgación de datos sensibles (Kaspersky, 2024; McAfee, 2023). La expansión de SpyLoan en América Latina y otros mercados muestra cómo el fraude financiero se articula con la vulnerabilidad económica y la baja alfabetización digital, sumando el daño financiero a la humillación y la presión social sobre la víctima (Kaspersky, 2023; McAfee, 2023).

El uso de cuentas mulas digitales constituye otra dimensión clave de los fraudes financieros asociados a plataformas de contenido. Las llamadas cuentas mulas son perfiles bancarios, financieros o móviles "de uso" que operan como intermediarios para el movimiento de capitales de origen ilícito, dividiendo grandes cantidades en múltiples transferencias de menor tamaño para dificultar la trazabilidad (BioCatch, 2024; FinReg360, 2025). Estas cuentas se captan a través de anuncios engañosos, ofertas de trabajo o servicios de recepción de dinero, donde se sugiere a los usuarios recibir pequeñas cantidades a cambio de una comisión, sin que estos comprendan plenamente que participan en esquemas de blanqueo (FinReg360, 2025; IDConline, 2025). En México, se ha documentado un incremento de más del 400 % en el uso de cuentas mulas entre 2021 y 2024, muchas de las cuales facilitan el blanqueo de dinero procedente de estafas digitales, tales como estafas románticas, robo de identidad, ataques de ransomware o campañas de ingeniería social (El País, 2025; Revista Seguridad, 2025). La normalización de este fenómeno indica que las plataformas digitales, en combinación con sistemas financieros y redes sociales, sirven como estructura operativa para el lavado de dinero a gran escala (BioCatch, 2024; FinReg360, 2025).

En paralelo, **la integración de estas prácticas en procesos de blanqueo de capitales** convierte a plataformas de contenido, redes sociales y servicios de pago en infraestructura crítica del crimen financiero. A través de perfiles de usuario legítimos, grupos de discusión, anuncios pagados y flujos de pago integrados, se fraccionan y disimulan fondos de origen delictivo, evitando el rastreo regulatorio y aprovechando la fragmentación de jurisdicciones y sistemas de detección (BioCatch, 2024; FinReg360, 2025). La conexión entre el delito financiero tradicional y la delincuencia digital se hace evidente cuando se observa que las cuentas mulas, utilizadas para el blanqueo, se originan en fraudes online, como estafas de inversión, robo de identidad o phishing masivo, lo que indica que el fraude en plataformas de contenido no solo genera daños directos, sino que se convierte en fuente de financiación para el crimen organizado (Revista Seguridad, 2025; IDConline, 2025).

En conjunto, los fraudes financieros asociados a plataformas de contenido digital se configuran como un ecosistema complejo en el que apps falsas, plataformas de inversión fraudulentas, modelos SpyLoan y cuentas mulas se articulan para captar patrimonio, datos sensibles y confianza de los usuarios (Kaspersky, 2025; Sin Embargo, 2025). La respuesta efectiva exige una combinación de controles de identidad digital, monitoreo de transacciones, detección de perfiles mula, conciencación de usuarios y regulación de plataformas financieras emergentes, con el objetivo de reducir la vulnerabilidad de los usuarios a esquemas que se disfrazan de oportunidades de inversión, crédito o servicios legítimos, pero cuya finalidad última es la explotación monetaria y la fragmentación de la trazabilidad financiera (BioCatch, 2024; FinReg360, 2025).

3.4. Fraudes basados en manipulación de información: desinformación, spam masivo, estafas románticas y deepfakes de personajes públicos.

Los fraudes basados en manipulación de información se configuran como un conjunto de estrategias que explotan la capacidad de la tecnología digital para alterar, exagerar o inventar contenidos con el fin de dirigir decisiones, emociones y comportamientos de usuarios en plataformas de contenido (ISMS Forum, 2026; IAON, 2026). A diferencia de los fraudes centrados en la usurpación técnica de cuentas o sistemas de pago, estos esquemas se dirigen a la percepción cognitiva del usuario, aprovechando la tendencia a confiar en fuentes familiares, en narrativas emocionales y en la apariencia de legitimidad visual o testimonial (ISMS Forum, 2026). En este contexto, se distinguen cuatro modalidades principales: desinformación, spam masivo, estafas románticas y deepfakes aplicados a personajes públicos, que se articulan a través de redes sociales, mensajería instantánea y plataformas de entretenimiento.

La desinformación constituye una de las manifestaciones más extendidas de este tipo de fraude, ya que genera contenido falso o tergiversado que se difunde a gran escala a través de redes sociales, foros, grupos de mensajería y plataformas de vídeo (ISMS Forum, 2026; Keepnet Labs, 2026). En 2024, se estimó que el volumen de deepfakes y contenido sintético generado por inteligencia artificial se había duplicado aproximadamente cada seis meses, situando el número de deepfakes compartidos en unos 500.000 archivos; para 2025, se proyectaba que este volumen alcanzara cerca de 8 millones de archivos, lo que refleja el carácter sistémico de la manipulación de información (ISMS Forum, 2026; IAON, 2026). La capacidad de las organizaciones criminales para generar noticias falsas, declaraciones de autoridades o videos de eventos alterados permite influir en decisiones financieras, políticas o de consumo, frecuentemente antes de que las plataformas puedan verificar su autenticidad (ISMS Forum, 2026; IAON, 2026).

Un segundo tipo de fraude basado en manipulación de información se centra en el **spam masivo y la difusión automatizada de mensajes engañosos a través de redes sociales,** correo electrónico y aplicaciones de mensajería (Kaspersky, 2025; McAfee, 2026). Estos mensajes suelen imitar comunicaciones oficiales, promociones de servicios legítimos o notificaciones de urgencia, induciendo al usuario a hacer clic en enlaces que redirigen a páginas de phishing, descargas de malware o plataformas de pago fraudulentas. En 2025, se observó un aumento significativo de campañas de spam por SMS que suplantaban servicios de peaje, bancos y administraciones públicas, aprovechando la urgencia digital y la normalización del pago por dispositivos móviles para generar riesgo de robo de datos, acceso no autorizado y manipulación de decisiones. Según McAfee, más de la mitad de

las personas encuestadas afirmaron haber sido estafadas o presionadas para enviar dinero o regalos a través de internet, lo que indica que el spam no solo se ha automatizado, sino que se ha vuelto altamente personalizado y orientado a la manipulación emocional (McAfee, 2026).

Un tercer grupo de fraude de manipulación de información se centra en **las estafas románticas**, que se han multiplicado gracias al uso de inteligencia artificial y plataformas de intercambio de contenidos (McAfee, 2026; NordVPN, 2026). Estas estafas se desarrollan principalmente en apps de citas y redes sociales, donde se crean perfiles falsos que utilizan fotografías robadas, historias elaboradas y técnicas de bombarding emocional para ganarse la confianza de la víctima. McAfee rindió un informe de 2026 en el que se reveló que más de la mitad de las personas encuestadas habían sido estafadas o presionadas para enviar dinero o regalos a alguien conocido por internet, y que el 26 % había sido abordada por un chatbot de IA que se hacía pasar por una persona real en una aplicación de citas o en redes sociales (McAfee, 2026). NordVPN reportó, además, que el spam de estafas románticas se concentra en plataformas como Snapchat e Instagram, donde se construyen relaciones más cercanas a través de conversaciones privadas y contenido compartido, facilitando el envío de enlaces peligrosos y la captación de datos bancarios (NordVPN, 2026). El uso de IA para generar perfiles y mensajes altamente persuasivos aumenta la eficacia de estas estafas, convirtiendo la manipulación emocional en un vector de fraude financiero y de daño psicológico extenso (McAfee, 2026; ISMS Forum, 2026).

Por último, **los deepfakes de suplantación de personajes públicos** representan una forma avanzada de manipulación de información, donde se utilizan videos **generados por inteligencia artificial** para simular declaraciones, imágenes o acciones de figuras reales y aumentar la credibilidad de un mensaje (ISMS Forum, 2026; IAON, 2026). En 2024, se documentaron varios casos de estafas masivas que utilizaban deepfakes de figuras como Elon Musk para promover inversiones fraudulentas en criptomonedas, generando pérdidas económicas millonarias (ISMS Forum, 2026). En 2025, el uso de deepfakes se extendió a otros ámbitos: campañas de desinformación electoral, promoción de productos falsos y ataques de ingeniería social que imitaban a políticos, celebridades o ejecutivos empresariales para legitimar esquemas de estafa (ISMS Forum, 2026; McAfee, 2026). La Fundación Europa para la Seguridad Digital (IAON) indicó que, en 2025, las pérdidas asociadas a fraudes financieros realizados mediante deepfakes superaron los 1.500 millones de dólares, mientras que el número de incidentes se duplicaba cada seis meses, lo que evidencia el rápido crecimiento de este tipo de fraude (IAON, 2026).

La combinación de estas modalidades —desinformación, spam, estafas románticas y deepfakes aplicados a personajes públicos— genera un escenario en el que la manipulación de información no solo alimenta fraudes individuales, sino que erosiona la confianza en la verdad, la reputación de instituciones y la estabilidad de procesos democráticos (ISMS Forum, 2026; IAON, 2026). La dificultad de detectar deepfakes, especialmente en el ámbito de la voz, junto con la facilidad de su difusión a través de redes sociales, aumenta la vulnerabilidad de usuarios de todas las edades, niños y adultos, a la hora de discernir entre información real y generada artificialmente. IA-ON advirtió que, entre 2024 y 2025, el uso de deepfakes en redes sociales superó los 500.000 archivos y se proyectaba que, en 2026, el 90 % del contenido digital podría ser generado por IA, lo que implica una transformación profunda en la forma en que los usuarios perciben la veracidad de lo que ven (ISMS Forum, 2026).

En conjunto, los fraudes basados en manipulación de información se configuran como un conjunto de prácticas diseñadas para aprovechar la capacidad de la IA y plataformas de contenido para alterar la percepción de la realidad y captar atención, confianza y dinero (ISMS Forum, 2026; IAON, 2026). La respuesta efectiva exige una combinación de educación digital, regulación de contenidos sintéticos y desarrollo de tecnologías de detección de deepfakes, así como la promoción de prácticas responsables de uso de IA por parte de los creadores de contenido y plataformas de distribución (ISMS Forum, 2026; IAON, 2026).

4. Técnicas, vectores y herramientas de ataque

Este capítulo describe las técnicas, vectores y herramientas empleados por los ciberdelincuentes para perpetrar fraudes en plataformas de contenido digital, desde la ingeniería social avanzada hasta la infraestructura criminal organizada, incluyendo las implicaciones estratégicas para España.

El fraude en plataformas digitales de contenido puede conceptualizarse como un fenómeno socio-técnico basado en la explotación intencionada de vulnerabilidades técnicas, organizativas y humanas. Los estándares internacionales del sector de telecomunicaciones definen el fraude como la explotación de debilidades en procesos o sistemas que generan pérdidas económicas u otros impactos relevantes (GSMA, 2025; CFCA, 2023).

Este fenómeno presenta una naturaleza adaptativa y evolutiva, en la que los atacantes combinan mecanismos tecnológicos y estrategias de manipulación humana para maximizar la eficacia de los ataques (Anderson, 2020; Levi & Smith, 2021). Tal como se analizará en el Capítulo 5, esta convergencia incrementa significativamente el impacto del fraude en el ecosistema digital.

Pero ¿qué es un **vector de ataque**? En ciberseguridad se denomina **vector de ataque** al método o canal concreto que un ciberdelincuente utiliza para explotar una vulnerabilidad, obtener acceso no autorizado o comprometer la confidencialidad, integridad o disponibilidad de un sistema, un servicio o un usuario dentro de un entorno digital. En el contexto de las plataformas de contenido, este concepto incluye, entre otros, el phishing, la distribución de malware a través de apps falsas, el uso de credenciales comprometidas o la explotación de configuraciones de seguridad deficientes en cuentas de usuario, sistemas de pago o publicidad digital.

En el sector de las telecomunicaciones, esta evolución se manifiesta en la aparición de vectores que operan parcialmente fuera del dominio de control del operador, alterando los modelos tradicionales de detección y mitigación. Un ejemplo paradigmático es el uso de SMS blasters, dispositivos capaces de enviar mensajes directamente a terminales móviles sin pasar por la red del operador, lo que permite eludir mecanismos clásicos de control como firewalls o sistemas de detección basados en CDR (GSMA FASG#34, 2026). Este fenómeno confirma la tendencia hacia la industrialización del fraude descrita en la literatura (Levi & Smith, 2021), pero introduce un elemento diferencial: la externalización del vector de ataque fuera de la infraestructura de red.

De forma complementaria, el fraude en servicios financieros digitales, particularmente en entornos de mobile money, ha experimentado un crecimiento significativo en escala, complejidad y sofisticación, posicionándose como uno de los principales riesgos operativos en ecosistemas digitales (GSMA FASG#33, 2025). A diferencia de modelos tradicionales de fraude en telecomunicaciones —centrados en bypass o explotación de red—, el fraude en pagos se caracteriza por su naturaleza híbrida, combinando ingeniería social, explotación de canales de comunicación y manipulación de flujos financieros. Este fenómeno se inserta en una tendencia más amplia donde las telecomunicaciones actúan como vector habilitador de fraude financiero, especialmente en esquemas de Authorised Push Payment (APP) (Ramsey, 2024).

Para comprender la complejidad de estos ataques resulta indispensable distinguir dos dimensiones fundamentales: **el método y el mecanismo**.

El **método** se refiere a la táctica psicológica o la narrativa que el atacante emplea para manipular a la víctima, mientras que el **mecanismo** constituye la herramienta técnica o la vulnerabilidad que se explota para ejecutar el fraude (Hadnagy, 2018). La mayoría de los ataques exitosos no son puramente técnicos ni puramente sociales: los atacantes combinan ambas dimensiones de forma sinérgica, empleando una táctica de ingeniería social para explotar una vulnerabilidad específica que les permita acceder a un mecanismo técnico. Esta interdependencia método-mecanismo es la razón por la que las defensas deben ser multicapa y abordar ambos planos de forma simultánea.

4.1 Ingeniería social avanzada

La ingeniería social constituye el vector de entrada predominante en los esquemas de fraude digital. Modalidades como phishing, smishing y vishing explotan sesgos cognitivos mediante la generación de urgencia, autoridad o confianza (Cialdini, 2009; Hadnagy, 2018). Los estándares sectoriales reconocen explícitamente la ingeniería social como una categoría relevante dentro de los esquemas de fraude, dada su capacidad para eludir controles técnicos mediante la manipulación del usuario (GSMA, 2025).

El vishing, en particular, ha experimentado un crecimiento significativo como vector de ataque. Esta técnica, que fusiona la interacción por voz con tácticas de phishing, aprovecha la confianza que los usuarios han depositado históricamente en las llamadas telefónicas como medio de comunicación directo y personal. Los discursos de ingeniería social que articulan estos ataques se apoyan en tres pilares fundamentales: el uso de la autoridad (suplantación de entidades oficiales o financieras), la generación de urgencia y miedo (amenazas de bloqueo de cuentas o acciones legales), y la manipulación emocional (apelación a vínculos personales o familiares). Estos pilares buscan una respuesta impulsiva y no analítica por parte de la víctima, siguiendo la misma lógica que los diseños de plataformas digitales que buscan captar la atención mediante estímulos emocionales (Cross, 2018).

La ejecución de la ingeniería social ha evolucionado hacia modelos altamente contextualizados. En campañas recientes, los mensajes fraudulentos suplantan entidades como bancos, organismos públicos o servicios digitales, generando narrativas diseñadas para inducir acciones inmediatas por parte del usuario (GSMA FASG#34, 2026). Por ejemplo, se han documentado mensajes que simulan alertas de seguridad o notificaciones administrativas con el objetivo de capturar credenciales o inducir transferencias fraudulentas. A diferencia de modelos tradicionales de phishing, estos ataques se benefician de canales percibidos como más confiables, como el SMS, y de su integración con vectores técnicos avanzados como los SMS blasters (véase Sección 4.5).

En este contexto, el crecimiento exponencial del phishing —con incrementos superiores al 800% desde 2019— evidencia la consolidación de modelos de fraude centrados en el usuario como punto de compromiso principal (GSMA FASG#33, 2025). Esta tendencia es especialmente relevante en entornos de mobile money, donde la ingeniería social se combina con la manipulación de flujos financieros para maximizar la monetización del ataque.

En el contexto de plataformas de contenido, estas técnicas se adaptan a patrones de consumo digital, integrándose en comunicaciones relacionadas con suscripciones, accesos premium o renovaciones de servicio. Los atacantes simulan notificaciones legítimas de servicios de streaming, plataformas de pago o proveedores de contenido para obtener credenciales de acceso. Como se discute en la Sección 1.2.2, este vector no solo genera pérdidas económicas, sino que afecta directamente a la confianza del usuario en el ecosistema digital.

4.2 Automatización, escala y economía del fraude

El fraude contemporáneo presenta características propias de sistemas industriales, incluyendo automatización, escalabilidad y optimización de costes. Este modelo se alinea con el concepto de economías criminales descrito por Anderson (2020), donde la eficiencia operativa y la especialización funcional permiten maximizar el retorno económico del fraude.

En el caso de los SMS blasters, se han observado capacidades de envío de hasta 100.000 mensajes por hora por dispositivo, así como campañas que alcanzan millones de víctimas en periodos reducidos (GSMA FASG#34, 2026). Estas capacidades, combinadas con costes de adquisición relativamente bajos, reducen significativamente las barreras de entrada para actores maliciosos. Asimismo, factores estructurales del sector telco, como la alta competencia y la diversificación de servicios, contribuyen a ampliar la superficie de ataque (GSMA FASG#34, 2026).

El fraude no solo se origina externamente; también puede involucrar actores internos o del ecosistema, como agentes o partners. Casos documentados muestran hasta un 35% de activaciones falsas, generación de comisiones fraudulentas y manipulación de incentivos comerciales. Este fenómeno pone de manifiesto que el fraude puede estar alineado con estructuras de incentivos mal diseñadas, lo que requiere controles en tiempo real sobre las operaciones de agentes (GSMA FASG#33, 2025). En este sentido, el fraude evoluciona desde un problema de seguridad hacia un problema de gobernanza de ecosistemas.

Tal como se desarrolló en la Sección 1.2.1, esta escalabilidad técnica se traduce en impactos económicos significativos.

4.3 Uso de inteligencia artificial y deepfakes

La incorporación de inteligencia artificial ha transformado significativamente la naturaleza del fraude digital. Aunque los marcos operativos tradicionales no contemplaban explícitamente estas tecnologías, los estándares sectoriales reconocen la evolución constante de las tipologías de fraude y la necesidad de adaptar los mecanismos de detección a nuevas amenazas (GSMA, 2025).

En la actualidad, la IA permite:

- **Automatización de campañas de fraude** a gran escala, incluyendo la generación masiva de llamadas fraudulentas (robocalls) con voces sintéticas personalizadas.

- **Personalización de ataques** en función del perfil de la víctima, mediante el análisis automatizado de datos públicos en redes sociales y plataformas digitales.

- **Generación de contenido sintético (deepfakes)** para suplantación de identidad, incluyendo la clonación de voz que permite reproducir con alta fidelidad la voz de familiares, directivos o figuras de autoridad.

- **Creación de sitios web y aplicaciones clonadas** con niveles de realismo sin precedentes, dificultando su distinción respecto de servicios legítimos.

La clonación de voz mediante IA constituye una evolución particularmente relevante para el fraude en plataformas de contenido. Con apenas unos segundos de audio original, las herramientas actuales permiten generar voces sintéticas capaces de engañar incluso a interlocutores cercanos a la víctima suplantada. Esta capacidad refuerza la tendencia hacia la industrialización del fraude (Levi & Smith, 2021) y amplía la superficie de ataque en un ecosistema digital donde la IA se integra progresivamente en publicidad, distribución de contenido y experiencia de usuario.

Tal como se ha desarrollado en la Sección 1.2.3, estas capacidades tienen implicaciones relevantes en términos de desinformación y manipulación social.

4.4 Malware, aplicaciones maliciosas y plataformas clonadas

El fraude técnico incorpora múltiples vectores catalogados en estándares sectoriales internacionales, que incluyen malware móvil, spoofing, phishing, manipulación de infraestructuras de red y explotación de vulnerabilidades en protocolos de señalización (GSMA, 2025). Los marcos de referencia del sector clasifican estas amenazas en categorías diferenciadas: fraude técnico (SIM cloning, PBX hacking, data charging bypass), fraude de suscripción (account takeover, eSIM fraud), fraude de distribución (dealer fraud, remote order fraud) y fraude de negocio (IRSF, content theft, wangiri).

Estos mecanismos permiten la ejecución de ataques orientados a:

- **Captura de credenciales** mediante phishing, pharming y técnicas de ingeniería social dirigidas tanto a usuarios finales como a intermediarios.

- **Acceso no autorizado a servicios**, incluyendo la explotación de centrales de conmutación (PBX hacking), la manipulación de perfiles en registros de localización (HLR tampering) y el clonado de tarjetas SIM.

- **Suplantación de identidad de llamada** (Caller ID Spoofing), que permite falsificar el número origen para generar confianza en la víctima, facilitando esquemas de vishing y fraude de callback.

- **Intercambio fraudulento de SIM** (SIM Swapping), que explota debilidades procesales en los operadores para transferir el número de la víctima a una tarjeta controlada por el atacante, permitiendo interceptar códigos de autenticación de dos factores.

- **Explotación de protocolos de señalización** (SS7), cuyo diseño basado en la confianza entre operadores permite la interceptación de comunicaciones y el desvío de mensajes sin interacción con la víctima.

- **Persistencia en sistemas comprometidos** mediante malware móvil diseñado para evadir mecanismos de detección convencionales.

En el ámbito de plataformas de contenido, estas técnicas se traducen en aplicaciones fraudulentas que simulan servicios legítimos de streaming, plataformas clonadas que replican la experiencia de usuario de servicios originales, y accesos ilícitos a contenidos protegidos. La existencia de múltiples mecanismos para lograr un mismo objetivo —como la interceptación de códigos 2FA, alcanzable tanto vía SIM swapping como vía explotación de SS7— subraya la diversidad de rutas de ataque disponibles y la necesidad de defensas multicapa (ENISA, 2023). Esta convergencia entre ciberfraude y modelos de distribución digital refuerza el impacto económico descrito en la Sección 1.2.1.

4.5 SMS blasters y estaciones base falsas

Los SMS blasters representan un vector de ataque especialmente relevante por su capacidad para operar fuera del plano de control de red. Se definen como estaciones base falsas portátiles que simulan redes legítimas para interactuar directamente con dispositivos móviles (GSMA FASG#34, 2026).

El funcionamiento del ataque sigue un proceso estructurado:

- Simulación de una estación base legítima (4G).

- Captura del dispositivo del usuario.

- Downgrade de la conexión a 2G (menos seguro).

- Inyección de mensajes SMS sin autenticación.

- Liberación del dispositivo hacia la red legítima.

Este ciclo puede completarse en intervalos de entre 15 segundos y 5 minutos (GSMA FASG#34, 2026).

A diferencia de otros vectores, estos ataques no generan registros en sistemas de facturación (CDR), eluden mecanismos de filtrado de red y presentan baja visibilidad para el operador. Como consecuencia, los modelos tradicionales de detección resultan insuficientes, lo que plantea desafíos significativos para la mitigación (véase Sección 5.3).

La evidencia empírica reciente confirma la adopción global de este vector. Se han documentado ataques en múltiples regiones, incluyendo Europa, Asia y Oriente Medio (GSMA FASG#34, 2026). Los casos analizados muestran patrones comunes: despliegue de dispositivos en entornos móviles (vehículos), uso de identidades falsas o suplantadas, integración con aplicaciones maliciosas y canales digitales, y coordinación mediante plataformas como Telegram. Se han identificado modelos organizativos descentralizados,

con células operativas de reducido tamaño pero alta movilidad, lo que dificulta su detección y desarticulación.

4.6 Fraude en mobile money y pagos digitales

Los sistemas de mobile money presentan características que los convierten en objetivos prioritarios: alta penetración, baja fricción en transacciones y dependencia de canales como SMS y voz. Se estima que un operador medio de mobile money puede perder aproximadamente 1,06 millones de dólares anuales debido al fraude, lo que impacta directamente en su rentabilidad y competitividad (GSMA FASG#33, 2025).

Más allá del impacto económico directo, el fraude genera efectos estructurales: erosión de confianza del usuario, incremento del churn (hasta un 33% en algunos casos) y reducción del volumen transaccional (hasta un 66%). Estos efectos refuerzan la idea de que el fraude no es únicamente un problema operativo, sino un factor estratégico que condiciona la sostenibilidad del modelo de negocio digital.

El fraude en mobile money sigue patrones operativos estructurados y repetibles. Un modelo típico incluye las siguientes fases:

1. Identificación de víctimas mediante análisis de datos.
2. Contacto a través de SMS o llamadas fraudulentas.
3. Manipulación mediante ingeniería social.
4. Transferencia de fondos hacia cuentas intermediarias (money mules).
5. Extracción rápida de fondos (cash-out).

Este flujo operativo refleja una cadena de valor criminal altamente optimizada, donde cada etapa puede ser ejecutada por actores distintos (GSMA FASG#33, 2025). Un elemento relevante es que aproximadamente el 40% de las víctimas reportan el fraude, lo que introduce una fuente crítica de señal para los sistemas de detección basados en comportamiento.

Este modelo confirma la convergencia entre fraude telco y fraude financiero, donde el canal de comunicación se convierte en el punto de entrada y el sistema de pagos en el vector de monetización.

4.7 Infraestructura criminal

El fraude digital se sustenta en un ecosistema estructurado que incluye múltiples categorías operativas. Las taxonomías sectoriales internacionales identifican al menos cuatro grandes ámbitos: fraude técnico, fraude de suscripción y pago, fraude de distribución y fraude de negocio (GSMA, 2025; CFCA, 2023). Esta clasificación refleja un alto grado de especialización funcional en la cadena de valor criminal.

Este modelo facilita la existencia de infraestructuras que permiten:

- **Comercialización de datos y credenciales** en mercados de la dark web, incluyendo listas de IMEIs fraudulentos, números B asociados a fraude premium y dominios irregulares.

- **Externalización de servicios de fraude** (Fraud-as-a-Service), donde organizaciones criminales ofrecen capacidades técnicas especializadas como SIM farms, plataformas de VoIP para robocalls o kits de phishing personalizados.

- **Monetización de accesos ilícitos** a través de esquemas de reventa de servicios, generación de tráfico artificial hacia números premium (International Revenue Share Fraud - IRSF) o arbitraje tarifario.

- **Redes de distribución de contenidos pirateados**, incluyendo servicios IPTV ilícitos, reventa de credenciales de plataformas de streaming y redistribución no autorizada de contenidos protegidos.

El fraude actual combina múltiples capas operativas: infraestructura física (dispositivos de ataque como SMS blasters y SIM boxes), infraestructura digital (plataformas, aplicaciones, dominios) e infraestructura organizativa (reclutamiento, monetización). Este modelo híbrido permite desacoplar las distintas fases del fraude, aumentando su resiliencia y escalabilidad (Levi & Smith, 2021; GSMA FASG#34, 2026). La externalización del vector de ataque fuera de la red introduce nuevos desafíos para su detección y control.

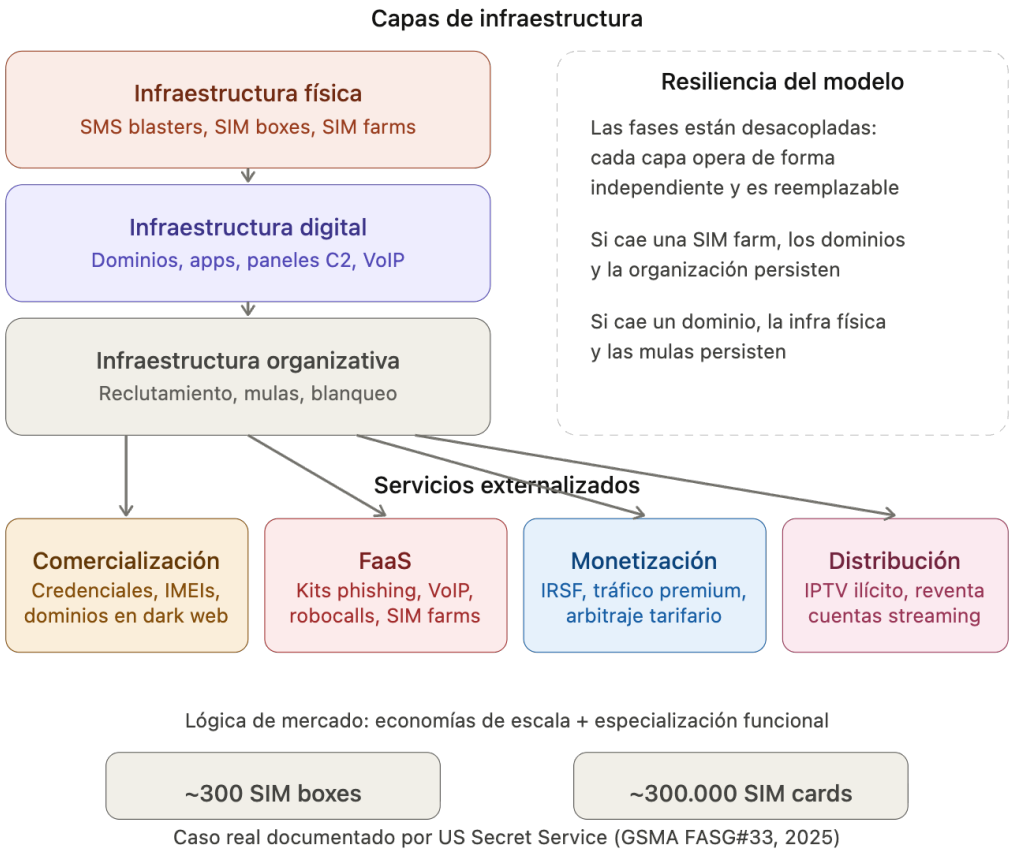

Figura 4 Ecosistema criminal Fraud-as-a-Service: tres capas de infraestructura (física, digital, organizativa) y cuatro categorías de servicios externalizados. Elaboración propia a partir de GSMA (2025), CFCA (2023) y Anderson (2020).

Un caso real de operación criminal detectada por el US Secret Service pone de manifiesto la escala de estas infraestructuras: se identificaron aproximadamente 300 SIM boxes y hasta 300.000 SIM cards, lo que evidencia las implicaciones para fraude, anonimización de comunicaciones y riesgos de seguridad nacional (GSMA FASG#33, 2025).

Como señalan estudios recientes, este modelo responde a una lógica de mercado con economías de escala y especialización funcional (Anderson, 2020). La comprensión de este ecosistema resulta clave para el diseño de respuestas regulatorias, como se discute en el Capítulo 5.

4.8 Fraude específico en servicios de TV y contenidos digitales

El sector de televisión y contenidos digitales presenta vectores de fraude específicos que merecen un tratamiento diferenciado. La transformación del sector audiovisual —con la migración acelerada desde modelos de TV de pago tradicional hacia plataformas de streaming y modelos híbridos de suscripción con publicidad— ha reconfigurado la superficie de ataque (Omdia, 2025).

Las principales tipologías de fraude en este ámbito incluyen:

- **Distribución no autorizada de contenidos**: la redistribución ilegal de contenidos propios o de distribución exclusiva a través de URLs, perfiles en redes sociales,

plataformas de compartición de archivos o servicios IPTV piratas. El impacto incluye tanto pérdidas económicas directas como daño reputacional derivado del uso fraudulento de marcas y derechos.

- **Reventa de servicios de TV**: la comercialización de suscripciones irregulares mediante anuncios en plataformas de compraventa, puntos de venta no autorizados o canales de distribución alternativos. Este tipo de fraude explota la compartición de credenciales y la creación de cuentas fraudulentas para generar accesos múltiples no autorizados.

- **Piratería de contenidos**: incluye la captura y redistribución de señales en directo, la eliminación de marcas de agua digitales (watermarking) y la explotación de vulnerabilidades en sistemas DRM (Digital Rights Management) para acceder a contenidos sin autorización.

La prevención del fraude en este ámbito requiere un enfoque transversal que involucre a las áreas técnicas, comerciales y de negocio del servicio de TV y contenidos de forma end-to-end. Según los marcos operativos del sector, las medidas preventivas incluyen la protección de contenidos mediante watermarking, el uso de herramientas digitales de detección automatizada, la monitorización de anuncios en plataformas de compraventa y la articulación de acciones legales coordinadas (GSMA, 2025). La eficacia de estas medidas depende de la capacidad de los operadores para establecer controles tanto primarios (integrados en el diseño del producto) como secundarios (implementados por áreas especializadas de prevención).

4.9 Implicaciones estratégicas para España

Las secciones precedentes han analizado el fraude digital en plataformas de contenido desde sus fundamentos técnicos, sus tipologías, sus vectores de ataque y sus consecuencias económicas y sociales. Esta sección sintetiza las implicaciones que ese análisis tiene para el contexto específico español, con el objetivo de dotar al trabajo de una dimensión operativa y de política pública que trascienda el diagnóstico y oriente la acción. España presenta una combinación de factores estructurales que hacen de su ecosistema digital un entorno de particular relevancia para el estudio del fraude: alta penetración de internet y servicios de streaming, un sector financiero y telco con presencia internacional significativa, y una economía con elevado volumen de transacciones digitales concentradas en pocas plataformas dominantes. Sobre este sustrato, el fraude digital opera con una eficacia y una escala que demandan respuestas estratégicas diferenciadas.

El fraude digital en plataformas de contenido ha transitado en España desde una problemática operativa —gestionada por equipos de seguridad sectoriales con respuestas reactivas— hacia un fenómeno estructural con implicaciones directas sobre la seguridad económica nacional, la confianza en la infraestructura digital y la estabilidad del ecosistema tecnológico. Los datos son elocuentes: 106.800 infracciones penales vinculadas a estafas informáticas en el primer trimestre de 2025 —equivalentes a 1.200 estafas diarias—, pérdidas directas que superan los 350 millones de euros anuales según el Banco de España, y un incremento interanual cercano al 40% que contrasta con la relativa estabilidad de otros tipos de delito (Europa Press, 2025; Visa España, 2025;

InfoBae, 2025b). Este contexto justifica un análisis de las implicaciones estratégicas a cuatro niveles.

Implicación 1 — Industrialización del fraude como economía paralela

La primera implicación estratégica deriva del proceso de industrialización del fraude documentado en el Capítulo 4. El fraude digital ha dejado de ser una actividad marginal o artesanal para convertirse en un sector económico ilícito con atributos propios de las industrias legítimas: división del trabajo, economías de escala, externalización de servicios (Fraud-as-a-Service), canales de distribución de productos (credenciales, accesos, malware) y modelos de negocio recurrentes con estructuras de ingresos predecibles. El informe de la Global Initiative Against Transnational Organized Crime (2026) estima que el fraude digital supera el billón de dólares anuales en ingresos ilícitos a escala global, y los datos del GSMA sitúan las pérdidas por fraude telco en más de 38.000 millones de dólares anuales, de los cuales una parte significativa impacta directamente en operadores con presencia relevante en España (GSMA, 2025; CFCA, 2023).

La implicación estratégica para España es doble. Por un lado, la respuesta al fraude no puede seguir el paradigma del incidente individual: requiere el reconocimiento explícito de que se enfrenta a una industria adversarial con recursos, resiliencia y capacidad de adaptación. Las organizaciones criminales detectadas por el Ministerio del Interior en los últimos años muestran estructuras jerárquicas con especialización funcional —reclutadores de mulas, operadores técnicos, lavadores de fondos— que no pueden desarticularse con enfoques puramente reactivos o sectoriales (Ministerio del Interior, 2026). Por otro lado, la industrialización del fraude genera externalidades económicas que van más allá del daño directo a las víctimas: incremento de los costes de gestión del riesgo para empresas y plataformas, primas de seguro más elevadas, inversión defensiva que desplaza recursos de la innovación, y deterioro de los indicadores de confianza que afectan a la adopción de servicios digitales y, en última instancia, a la competitividad del tejido empresarial español en el mercado digital europeo.

Implicación 2 — Convergencia intersectorial como vector de riesgo sistémico

La segunda implicación estratégica emerge de la convergencia entre sectores que caracteriza el fraude digital contemporáneo. El análisis de los vectores de ataque (Capítulo 4) y de la infraestructura criminal (Sección 4.7) evidencia que los ataques exitosos no se limitan a un único sector regulatorio: utilizan la red de telecomunicaciones como canal de captación (smishing, vishing, SMS blasters), las plataformas de contenido digital como vector de distribución y compromiso (credential stuffing, account takeover, malware en apps), y el sistema financiero como mecanismo de monetización y blanqueo (transferencias a cuentas mulas, conversión a criptomonedas). Esta arquitectura de ataque distribuida implica que la detección y mitigación efectiva requiere correlacionar señales de los tres sectores, algo que los marcos regulatorios y los sistemas de información actuales en España no están diseñados para hacer.

El sistema de supervisión y regulación español opera con silos sectoriales relativamente bien definidos: la CNMC supervisa las telecomunicaciones, el Banco de España y la CNMV el sistema financiero, y la AEPD la protección de datos. Sin embargo, ninguno de estos organismos dispone de una visión transversal de los flujos de fraude que atraviesan los tres sectores simultáneamente. La Directiva NIS2, en proceso de

transposición, introduce algunos mecanismos de coordinación, pero su alcance sobre el fraude en plataformas de contenido es limitado y requiere un desarrollo específico. Esta fragmentación regulatoria es, en sí misma, una vulnerabilidad sistémica que los actores del fraude explotan activamente: la discontinuidad en la supervisión entre el operador telco, la plataforma y la entidad financiera genera espacios de actuación que ningún regulador individual tiene incentivos suficientes para cubrir.

Implicación 3 — Erosión de la confianza digital como riesgo macroeconómico

La tercera implicación estratégica opera a un nivel más difícilmente cuantificable pero potencialmente más relevante para el modelo de desarrollo económico español: el impacto del fraude digital sobre la confianza en la economía digital como plataforma de crecimiento. España aspira a posicionarse como hub digital europeo, con la agenda España Digital 2026 y los fondos PERTE para la transformación digital como vectores de inversión. Este posicionamiento depende críticamente de que los ciudadanos, las empresas y los inversores confíen en la seguridad y fiabilidad de la infraestructura digital española. El fraude erosiona esa confianza por múltiples vías.

La adopción de servicios de banca digital y comercio electrónico muestra resistencias especialmente pronunciadas en segmentos de mayor edad o menor alfabetización digital, precisamente los grupos más expuestos al fraude (Digital Innovation News, 2026; Management Society, 2025). Las encuestas de Visa España (2025) señalan que el 44% de los españoles declara tener dificultades para distinguir contenido auténtico del generado por IA, lo que refleja una vulnerabilidad cognitiva estructural ante los vectores de fraude basados en deepfakes y contenido sintético. A nivel empresarial, más de la mitad de las grandes empresas españolas ha sufrido los efectos de desinformación vinculada a su marca en 2025, con impacto directo sobre valoraciones bursátiles, relaciones con clientes y costes de comunicación corporativa (El Debate, 2024). Esta confluencia de efectos sobre la confianza ciudadana y empresarial configura un riesgo macroeconómico que trasciende el daño operativo del fraude individual y que justifica su tratamiento como asunto de política económica, no solo de seguridad.

Implicación 4 — La inteligencia artificial como asimetría estructural

La cuarta implicación estratégica es la más reciente en términos de emergencia y la que plantea los retos más complejos para la respuesta institucional: el uso de la inteligencia artificial como multiplicador asimétrico del riesgo de fraude. El análisis de la Sección 4.3 documenta la aceleración sin precedentes del fraude basado en IA: la duplicación del volumen de deepfakes cada seis meses, el uso de modelos de lenguaje para la generación de correos de phishing hiperpersonalizados, la automatización masiva de campañas de smishing y la creación de perfiles falsos con comportamiento humano convincente en redes sociales (ISMS Forum, 2026; IAON, 2026; Kaspersky, 2026).

La asimetría que introduce la IA en el ecosistema del fraude es estructural, no coyuntural: la misma tecnología que los defensores utilizan para entrenar modelos de detección de anomalías está disponible para los atacantes a un coste marginal decreciente y en un contexto regulatorio que aún no ha alcanzado madurez operativa. El AI Act europeo, aprobado en 2024 con implantación progresiva hasta 2027, establece obligaciones de transparencia y robustez para los sistemas de IA de alto riesgo (Unión Europea, 2024), pero su aplicación específica al fraude en plataformas de contenido requiere un desarrollo reglamentario que los reguladores nacionales deben anticipar, no esperar. España, como

país con capacidad de influencia regulatoria en el proceso de desarrollo de actos delegados del AI Act, tiene una oportunidad de posicionamiento activo en la definición de estándares que afectarán directamente a la eficacia de sus sistemas de defensa frente al fraude habilitado por IA. La coordinación entre INCIBE, la Agencia Española de Supervisión de la Inteligencia Artificial (AESIA) y el Ministerio de Transformación Digital es, en este sentido, una necesidad operativa de primer orden.

5. Marco legal y regulatorio aplicable

El Capítulo 5 aborda el marco legal y regulatorio aplicable al fraude digital en el ecosistema de plataformas de contenido, analizando tanto la normativa existente como la evolución de los estándares internacionales y la responsabilidad de los distintos actores. Frente a un fenómeno cada vez más complejo, globalizado y tecnológicamente sofisticado, los sistemas jurídicos se ven obligados a articular respuestas multinivel, que articulen legislación penal, regulación de servicios digitales, protección de datos y normas de seguridad financiera. En este contexto, se estudian no solo los instrumentos normativos, sino también la responsabilidad de plataformas y operadores telco, la evolución hacia modelos de prevención en tiempo real y los mecanismos de cooperación internacional que condicionan la efectividad de las políticas de ciberseguridad.

5.1 Marcos normativos

El fraude digital se aborda mediante un conjunto de marcos normativos que incluyen legislación penal, regulación de servicios digitales, protección de datos y propiedad intelectual. Estos marcos operan en un contexto de gobernanza multinivel donde coexisten estándares globales, regulación supranacional y legislación nacional.

A nivel internacional, los estándares sectoriales proporcionan definiciones comunes, taxonomías de fraude y mecanismos de mitigación que facilitan la coordinación entre actores del ecosistema digital. El Fraud Manual de la GSMA, actualmente en su versión 22.0, constituye el marco de referencia más amplio del sector de telecomunicaciones, catalogando más de cuarenta tipologías de fraude organizadas en categorías funcionales (GSMA, 2025). La Communications Fraud Control Association (CFCA) complementa este marco con encuestas periódicas que cuantifican las pérdidas globales por fraude y los foros sectoriales como TMForum proporcionan estándares para la gestión del aseguramiento de ingresos.

En el contexto europeo, el Digital Services Act (DSA) refuerza las obligaciones de las plataformas digitales en materia de transparencia, moderación de contenidos ilícitos y protección de usuarios frente a prácticas fraudulentas. El Reglamento General de Protección de Datos (RGPD) establece principios de transparencia, proporcionalidad y legitimación que condicionan los tratamientos de datos necesarios para la detección de fraude, exigiendo que las medidas de prevención respeten los derechos y libertades de los interesados.

En materia de autenticación de comunicaciones, iniciativas como el protocolo STIR/SHAKEN en Estados Unidos establecen mecanismos de firma digital para verificar la autenticidad del identificador de llamada, constituyendo una respuesta regulatoria directa al problema del Caller ID Spoofing. En el ámbito europeo, se están desarrollando marcos regulatorios equivalentes para combatir las estafas por suplantación telefónica.

El Reglamento Europeo de Inteligencia Artificial (Reglamento (UE) 2024/1689), con plena aplicabilidad a partir de agosto de 2026, introduce un marco regulatorio estratificado por niveles de riesgo que tiene implicaciones directas para el fraude en plataformas de contenido en tres dimensiones.

En la primera dimensión, la de transparencia, el artículo 50 establece que los sistemas de IA diseñados para interactuar directamente con personas físicas deben informar al usuario de que está interactuando con un sistema de IA, salvo que resulte evidente por las circunstancias. Esta obligación afecta directamente a los chatbots y asistentes virtuales desplegados en plataformas de contenido — como Aura de Telefónica o los asistentes de atención al cliente de las principales plataformas de streaming — y tiene una conexión directa con el fraude: los atacantes que suplantan estos asistentes mediante chatbots fraudulentos operan en un espacio donde la obligación de transparencia del AI Act podría convertirse en un criterio de diferenciación entre servicio legítimo y vector de ataque. Asimismo, el artículo 50.4 exige el etiquetado de contenido generado por IA cuando se trate de imagen, audio o vídeo que constituya un deep fake, una disposición cuya aplicabilidad a los deepfakes utilizados en estafas de inversión (Caso B del Capítulo 8) dependerá de la capacidad de las plataformas para implementar mecanismos de detección y etiquetado antes de la distribución del contenido fraudulento a través de sus redes de publicidad.

En la segunda dimensión, la de clasificación de riesgo, los sistemas de IA utilizados para la evaluación de la solvencia crediticia y la puntuación de crédito se clasifican como de alto riesgo (Anexo III, punto 5.b), lo que afecta directamente a los modelos de detección de fraude desplegados por entidades financieras cuando operan sobre datos de usuarios de plataformas de contenido — por ejemplo, los modelos de detección de cuentas mula que BioCatch (2025) describe como behavioral biometrics. Los proveedores de estos sistemas de alto riesgo quedarán sujetos a obligaciones de gestión de riesgos, gobernanza de datos, transparencia, supervisión humana y robustez técnica que condicionarán el diseño de los sistemas antifraude del Capítulo 6.

En la tercera dimensión, la de prácticas prohibidas, el artículo 5 prohíbe los sistemas de IA que desplieguen técnicas subliminales, manipulativas o engañosas que distorsionen de manera sustancial el comportamiento de una persona. Esta prohibición, concebida para proteger la autonomía del usuario, establece un principio regulatorio que conecta directamente con las técnicas de ingeniería social avanzada documentadas en la Sección 4.1: cuando un atacante utiliza IA generativa para crear deepfakes o chatbots que manipulan al usuario para que realice transferencias voluntarias (Caso B), la línea entre la técnica de fraude y la práctica prohibida por el AI Act se difumina, abriendo un espacio para la aplicación de sanciones regulatorias además de penales.

La Directiva NIS2 (Directiva (UE) 2022/2555), cuyo plazo de transposición venció en octubre de 2024, amplía significativamente el ámbito de aplicación de la normativa de ciberseguridad respecto a su predecesora. En el contexto del fraude en plataformas de contenido, la NIS2 es relevante en tres aspectos. Primero, la ampliación del perímetro de entidades obligadas: la directiva incluye por primera vez a proveedores de plataformas de redes sociales, servicios de computación en la nube y servicios de centros de datos como entidades esenciales o importantes, lo que somete a muchas plataformas de contenido digital a obligaciones de gestión de riesgos de ciberseguridad que antes no les eran exigibles. Segundo, las obligaciones de notificación de incidentes: el artículo 23 establece un plazo de 24 horas para la alerta temprana y 72 horas para la notificación completa al CSIRT de referencia, lo que implica que un incidente de fraude masivo como el descrito en el Caso A del Capítulo 8 activaría obligaciones formales de reporte cuyo incumplimiento conlleva sanciones. Tercero, la seguridad de la cadena de suministro: el artículo 21.2.d exige que las entidades obligadas evalúen los riesgos de

ciberseguridad de su cadena de proveedores, lo que obliga a las plataformas de contenido a verificar la seguridad de las pasarelas de pago, servicios de publicidad y proveedores de infraestructura que integran — precisamente las interfaces que configuran la superficie de ataque descrita en la Sección 2.2.

En España, la transposición de la NIS2 se ha materializado mediante la actualización del marco del Esquema Nacional de Seguridad (CCN-CERT, 2022) y la adaptación de las competencias de INCIBE como CSIRT de referencia para el sector privado. Sin embargo, la implementación efectiva de las obligaciones de la NIS2 en el ecosistema de plataformas de contenido presenta desafíos específicos: la fragmentación del sector entre operadores globales (cuya entidad de supervisión principal puede estar en otro Estado miembro) y actores nacionales, la heterogeneidad de los niveles de madurez en seguridad, y la dificultad de aplicar las obligaciones de gestión de riesgos de la cadena de suministro cuando la cadena incluye actores en jurisdicciones no europeas donde opera la infraestructura criminal documentada en la Sección 4.7.

Sin embargo, los sistemas tradicionales de detección de fraude presentan limitaciones significativas. Estos sistemas se basan en la monitorización del tráfico dentro de la red (e.g., CDR, signaling), pero resultan insuficientes frente a vectores que operan fuera de la red, como los SMS blasters descritos en la Sección 4.5 (GSMA FASG#34, 2026). Asimismo, los mecanismos de detección individuales presentan altas tasas de falsos positivos, lo que limita su efectividad cuando se aplican de forma aislada.

El fraude en pagos ha alcanzado un nivel de relevancia tal que se ha situado en el centro de la agenda regulatoria global (Ramsey, 2024). Particularmente, el fraude APP ha adquirido visibilidad debido a su impacto directo sobre consumidores, lo que ha impulsado propuestas regulatorias orientadas a reforzar la protección del usuario, aumentar la responsabilidad de los actores del ecosistema y mejorar el intercambio de información. En Europa, iniciativas como PSD2 y los desarrollos hacia PSD3 reflejan esta tendencia, introduciendo nuevas obligaciones para proveedores de servicios de pago y, potencialmente, para operadores telco (Ramsey, 2024).

5.2 Responsabilidad de las plataformas y debate sobre los operadores telco

Las plataformas digitales desempeñan un papel central en la prevención del fraude. Su responsabilidad se articula en torno a la implementación de mecanismos de:

- **Detección de actividades fraudulentas**, mediante soluciones antifraude que incorporen reglas de detección, modelos predictivos basados en aprendizaje automático y análisis de patrones de comportamiento anómalos (análisis estadísticos, series temporales, minería de datos).

- **Moderación de contenidos engañosos**, incluyendo la identificación y retirada de plataformas clonadas, aplicaciones fraudulentas y anuncios de servicios irregulares.

- **Protección de usuarios**, mediante mecanismos robustos de autenticación (preferiblemente multifactor basado en aplicaciones, no en SMS), cifrado de comunicaciones y transparencia en el tratamiento de datos.

Los marcos operativos del sector establecen que la prevención del fraude es una responsabilidad transversal de la organización, no limitada a un área funcional específica. Este enfoque requiere que las áreas de prevención de fraude participen en el diseño de nuevos productos y servicios desde sus fases iniciales, y que las áreas custodias de información proporcionen la colaboración necesaria para la investigación y resolución de eventos de fraude. La efectividad de la prevención depende de la capacidad de establecer controles primarios (integrados en el producto o proceso por sus propietarios) y controles secundarios (implementados por áreas especializadas de seguridad) de forma coordinada.

Uno de los elementos más controvertidos en la agenda regulatoria actual es la posible asignación de responsabilidad financiera a los operadores de telecomunicaciones en casos de fraude por suplantación. Algunas propuestas regulatorias plantean que los operadores deberían verificar el origen de comunicaciones, bloquear números fraudulentos, prevenir la creación de infraestructuras maliciosas y asumir parte del coste de compensación a víctimas. Sin embargo, el sector telco argumenta que no tiene visibilidad sobre la transacción financiera, no controla todas las fases del fraude y que la imposición de responsabilidad podría desincentivar la inversión en prevención. Este debate refleja un problema estructural: la asimetría de control y responsabilidad entre telecomunicaciones y servicios financieros (GSMA FASG#33, 2025; Ramsey, 2024).

Frente a modelos de responsabilidad unilateral, emergen enfoques de responsabilidad compartida, donde bancos, operadores telco y plataformas tecnológicas comparten obligaciones en la prevención y mitigación del fraude. Ejemplos como el marco regulatorio de Singapur introducen mecanismos estructurados para asignar responsabilidades en función del punto de fallo dentro de la cadena de fraude (Ramsey, 2024). Estos enfoques reconocen que el fraude es un fenómeno sistémico que no puede ser abordado desde un único actor.

Este enfoque requiere modelos integrados que combinen tecnología, procesos y gobernanza (Gillespie, 2018). Asimismo, la naturaleza transnacional del fraude exige mecanismos de cooperación internacional, en línea con la necesidad de abordar el ecosistema criminal descrito en la Sección 4.7.

La interacción entre el DSA y los mecanismos de certificación de seguridad constituye un ámbito regulatorio en desarrollo que tiene implicaciones directas para las recomendaciones de este trabajo. El DSA, en su artículo 34, obliga a las plataformas en línea de muy gran tamaño (VLOPs, con más de 45 millones de usuarios activos mensuales en la UE) a realizar evaluaciones de riesgos sistémicos que incluyan, entre otros, los riesgos de difusión de contenido ilegal y los efectos negativos sobre la protección de los consumidores. El artículo 35 exige además que estas plataformas adopten medidas de mitigación razonables, proporcionadas y eficaces frente a los riesgos identificados. En este marco, la propuesta de sello de confianza antifraude formulada en la Recomendación R9 de este trabajo podría articularse como un instrumento de cumplimiento voluntario que acredite la adopción de medidas de mitigación específicas frente al riesgo de fraude — una categoría de riesgo sistémico que el DSA reconoce implícitamente pero que las evaluaciones de riesgo publicadas

hasta la fecha (Comisión Europea, 2024) han abordado de forma tangencial frente a la desinformación y el contenido ilegal.

La viabilidad de esta articulación se apoya en dos precedentes regulatorios. En primer lugar, el propio DSA (artículo 45) contempla la elaboración de códigos de conducta y la posibilidad de que la Comisión solicite la creación de esquemas de certificación sectoriales como complemento a las obligaciones directas. En segundo lugar, el Reglamento de Ciberseguridad (Reglamento (UE) 2019/881) establece el marco europeo para esquemas de certificación de ciberseguridad, con ENISA como organismo de referencia, un modelo organizativo directamente aplicable al sello propuesto. La conexión entre ambos instrumentos — un sello de confianza antifraude que opere como presunción de cumplimiento parcial de las obligaciones del DSA en materia de mitigación de riesgos sistémicos, certificado bajo el marco del Reglamento de Ciberseguridad — proporcionaría el anclaje regulatorio necesario para que la propuesta de la Recomendación R9 trascienda el ámbito de la voluntariedad y genere incentivos económicos tangibles para las plataformas.

5.3 Prevención en tiempo real y nuevos enfoques de detección

Uno de los principales aprendizajes derivados del análisis del fraude contemporáneo es la transición desde modelos reactivos hacia modelos de prevención en tiempo real. El uso combinado de datos telco (e.g., comportamiento de llamadas, IMEI, Cell ID) y datos financieros permite bloquear transacciones fraudulentas en el momento de ejecución, detectar patrones de comportamiento anómalos y aplicar controles dinámicos tras eventos críticos como un SIM swap (GSMA FASG#33, 2025).

Este enfoque representa un cambio fundamental respecto a modelos tradicionales basados en análisis posterior (post-event detection), alineándose con arquitecturas de decisión en tiempo real. Además, introduce nuevas capacidades operativas, como throttling de transacciones, bloqueo de dispositivos (IMEI) y suspensión automática de cuentas.

En el ámbito de la detección de vectores que operan fuera de la red, los enfoques emergentes se basan en la correlación de múltiples fuentes de información, incluyendo anomalías en la red radio, patrones de comportamiento de usuarios, análisis geoespacial y despliegue de honeypots (GSMA FASG#34, 2026). Estos modelos reflejan una transición hacia sistemas de detección multi-capa, necesarios para abordar la complejidad del fraude actual.

A nivel arquitectónico, la evolución hacia modelos de detección avanzados se enmarca en frameworks como el Threat Register (FS.30) y Baseline Controls (FS.31) promovidos por el GSMA Fraud and Security Architecture Group, que incluyen controles sobre APIs, identidad y credenciales, así como la evolución hacia arquitecturas de confianza descentralizada (GSMA FSAG, 2025).

5.4 Cooperación internacional y estándares sectoriales

La respuesta eficaz al fraude en plataformas digitales exige una cooperación sin precedentes entre actores públicos y privados a escala internacional. Los mecanismos de coordinación actuales incluyen:

- **Compartición de información sobre eventos de fraude relevantes entre operadores y plataformas**, incluyendo listas de identificadores fraudulentos (IMEIs, números de destino, dominios irregulares) y patrones de ataque emergentes.

- **Coordinación con agencias gubernamentales e instituciones internacionales** como Europol, Interpol y agencias nacionales de ciberseguridad (ENISA en Europa, INCIBE en España, FTC en Estados Unidos) para la investigación y persecución de redes criminales organizadas.

- **Participación en foros sectoriales** (GSMA, CFCA, TMForum, ACFE) que facilitan el intercambio de mejores prácticas, la consolidación de inteligencia sobre amenazas y la estandarización de métricas de impacto.

- **Utilización de fuentes de inteligencia abiertas** (OSINT) y servicios de monitorización de la dark web para anticipar nuevos vectores de fraude y detectar la comercialización de credenciales y accesos ilícitos.

Dada la naturaleza transnacional del fraude, la cooperación entre operadores, reguladores y organismos internacionales resulta esencial. Iniciativas sectoriales como las promovidas en el marco del GSMA Fraud and Security Group facilitan el intercambio de inteligencia y la estandarización de prácticas (GSMA FASG#34, 2026). La estructura distribuida del fraude, descrita en la Sección 4.7, requiere respuestas igualmente coordinadas y globales.

A nivel global, se observa una convergencia en las estrategias regulatorias, incluyendo registros de identidad (Sender ID, SIM registration), uso de inteligencia artificial para detección, plataformas de intercambio de información y campañas de concienciación. Casos en países como India, España o Reino Unido muestran la adopción de medidas específicas para combatir fraude en comunicaciones y pagos (GSMA FASG#33, 2025). Este movimiento hacia la convergencia regulatoria sugiere la consolidación de un nuevo paradigma donde la prevención del fraude se convierte en una responsabilidad transversal del ecosistema digital.

En el ámbito específico de los servicios de pago, la evolución de PSD2 hacia PSD3 y el nuevo Reglamento de Servicios de Pago (PSR) introducen cambios significativos para la distribución de responsabilidades en el fraude. La propuesta de la Comisión Europea, actualmente en fase de negociación, contempla tres elementos con impacto directo sobre el ecosistema de plataformas de contenido: la extensión de las obligaciones de verificación del beneficiario (Confirmation of Payee / Verification of Payee) a todas las transferencias instantáneas en la zona euro, lo que dotaría al sistema financiero de una herramienta preventiva clave frente al fraude APP documentado en la Sección 4.1; la posible inclusión de los proveedores de servicios de comunicaciones electrónicas en el ámbito de las obligaciones de prevención del fraude, reflejando la tendencia hacia modelos de responsabilidad compartida como el implementado en Singapur (Ramsey, 2024); y el refuerzo de los mecanismos de intercambio de datos de fraude entre proveedores de servicios de pago, bajo un marco de protección de datos compatible con el RGPD, que constituiría el equivalente financiero de la Plataforma Nacional de Inteligencia Antifraude (PNIA) propuesta en la Recomendación R1.

La convergencia entre DSA, NIS2, AI Act y PSD3 configura un marco regulatorio que, aunque no diseñado específicamente para abordar el fraude en plataformas de contenido como fenómeno unitario, proporciona los instrumentos jurídicos necesarios para articular una respuesta coordinada. El reto para el legislador español y para INCIBE como organismo de referencia reside en conectar estos instrumentos en una estrategia coherente que evite la fragmentación regulatoria que el propio fraude explota como ventaja operativa.

La consolidación de información derivada de eventos de fraude que afectan a múltiples operadores o países permite generar conocimiento anticipativo y estratégico que refuerza la capacidad de respuesta del ecosistema en su conjunto. La eficacia de estos mecanismos de cooperación depende del establecimiento de canales de comunicación estandarizados, protocolos de confidencialidad adecuados y sistemas internos de información que garanticen la integridad y trazabilidad de los datos compartidos.

5.5 Implicaciones estratégicas para el sector telco

El análisis conjunto del fraude técnico, el fraude en pagos y los vectores emergentes como los SMS blasters evidencia una transformación estructural del rol de los operadores de telecomunicaciones. Tradicionalmente centrados en la protección de ingresos propios, los operadores se enfrentan ahora a la necesidad de evolucionar hacia modelos orientados a la protección del cliente y del ecosistema (GSMA FASG#33, 2025).

Este cambio implica inversión en capacidades de análisis en tiempo real, integración con actores financieros, redefinición de responsabilidades y alineación con marcos regulatorios emergentes. En este contexto, el fraude deja de ser un problema operativo aislado para convertirse en un elemento central de la estrategia digital.

Solo a través de un enfoque holístico que integre protección tecnológica, educación de usuarios, regulación adaptativa y cooperación internacional será posible mitigar de forma eficaz una amenaza que evoluciona al ritmo de la innovación digital.

6. Mecanismos de detección, prevención e investigación del fraude en plataformas de contenidos digitales

El fraude en plataformas de contenidos digitales no puede contenerse mediante controles puntuales aplicados en un único punto de la cadena de valor. Su naturaleza distribuida, adaptativa y multi-actor —documentada en los Capítulos 3, 4 y 5 de este trabajo— exige una arquitectura de defensa organizada en capas que opera simultáneamente en el plano técnico, el organizativo y el interinstitucional. Este capítulo analiza los componentes de esa arquitectura en cuatro dimensiones: los controles técnicos en plataformas digitales (autenticación, detección de anomalías, protección perimetral e inteligencia artificial), la investigación digital y el análisis forense, la colaboración público-privada, y los principios de integración que articulan estas dimensiones en un sistema coherente de defensa. El análisis parte de los mecanismos existentes y sus limitaciones documentadas, para derivar implicaciones sobre las capacidades que los actores del ecosistema español necesitan desarrollar.

6.1. Introducción: la defensa en capas como principio organizador

El principio de defensa en profundidad —defense in depth— establece que ningún control técnico aislado es suficiente para contener un atacante determinado: la seguridad efectiva emerge de la superposición de múltiples controles independientes, de forma que el fallo de cualquiera de ellos no compromete el sistema en su conjunto (Anderson, 2020). En el contexto del fraude en plataformas de contenido, este principio adquiere una dimensión adicional: las capas de defensa no pueden limitarse al perímetro técnico de una sola plataforma, sino que deben extenderse a través de los sectores que forman la cadena del ataque —telecomunicaciones, plataforma digital y sistema financiero—, porque es precisamente en las discontinuidades entre esos sectores donde el fraude más sofisticado encuentra su espacio de actuación.

Esta sección organiza los mecanismos de defensa en cuatro capas funcionales. La primera es la capa de autenticación y control de acceso, cuyo objetivo es verificar la identidad del usuario y detectar intentos de suplantación en el punto de entrada. La segunda es la capa de detección de anomalías y comportamiento, que opera una vez que el usuario está dentro del sistema y busca identificar comportamientos inconsistentes con el perfil legítimo o con patrones de uso normal. La tercera es la capa de protección perimetral y de aplicaciones, orientada a filtrar tráfico malicioso e impedir la automatización no autorizada. La cuarta es la capa de investigación y respuesta, que actúa cuando el fraude ya se ha producido para limitar el daño, recuperar evidencia y sustentar la persecución. A estas cuatro capas técnicas se añade la dimensión transversal de la colaboración público-privada, que determina en qué medida los controles de cada actor se suman de forma coherente o permanecen fragmentados en silos que el fraude puede explotar.

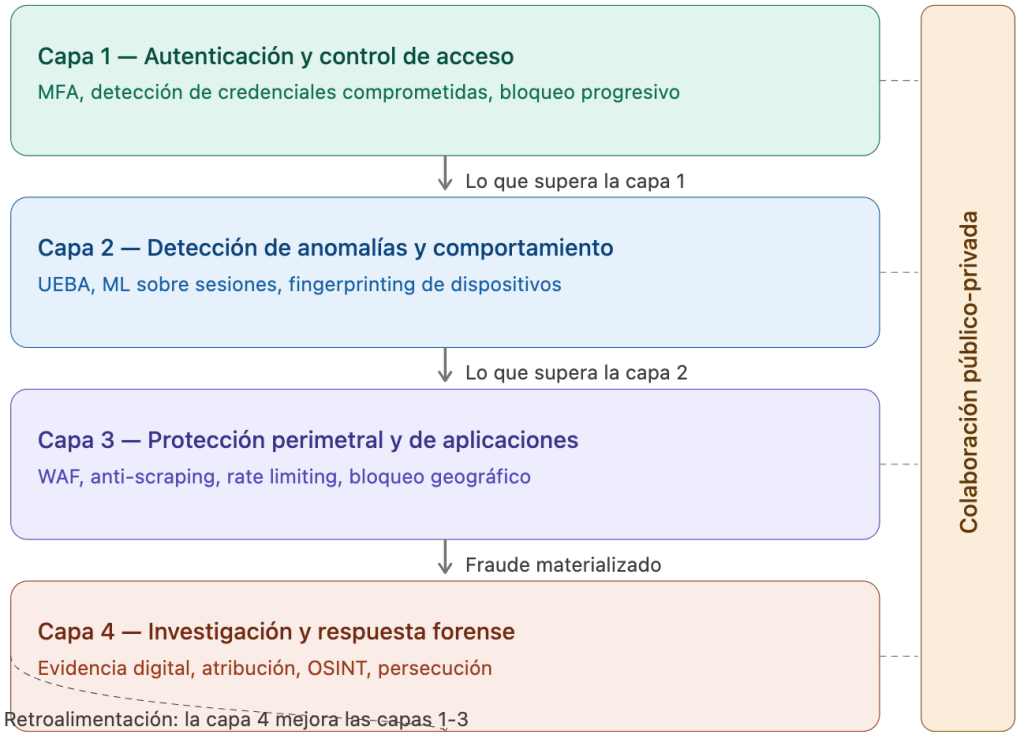

Figura 5 Arquitectura de defensa en profundidad para plataformas de contenido digital: cuatro capas funcionales con retroalimentación y dimensión transversal de colaboración público-privada. Elaboración propia a partir de Anderson (2020) y GSMA (2025).

6.2. Controles técnicos en plataformas digitales

6.2.1. Autenticación reforzada y control de acceso

La autenticación es la primera línea de defensa de cualquier plataforma digital y, al mismo tiempo, el control que los atacantes han aprendido mejor a circunvalar. La evolución desde la autenticación basada exclusivamente en contraseña hacia esquemas de múltiples factores ha reducido significativamente la eficacia de los ataques de credential stuffing y fuerza bruta, pero ha generado una asimetría entre plataformas con diferente nivel de madurez: en aquellas donde la MFA es opcional, los atacantes concentran sus esfuerzos precisamente en las cuentas sin segundo factor, que continúan representando la mayoría en muchos servicios de contenido digital (Phua et al., 2010). Google publicó en 2023 que la activación de MFA en sus servicios redujo el riesgo de account takeover en un 99,9%, pero también documentó que la tasa de adopción voluntaria por parte de los usuarios en servicios donde la MFA no es obligatoria se mantiene por debajo del 35% (Google, 2023). Esta brecha entre la eficacia demostrada del control y su tasa de adopción real es uno de los argumentos más sólidos a favor de la obligatoriedad regulatoria descrita en la Recomendación R6 en la Sección 7.2.

La autenticación adaptativa basada en riesgo representa la evolución más relevante en este espacio durante los últimos cinco años. A diferencia de la MFA estática —que aplica el mismo nivel de verificación independientemente del contexto—, la autenticación adaptativa calibra el nivel de fricción requerido en función de variables de riesgo como

la geolocalización del acceso, el dispositivo utilizado, la hora del día, el historial de comportamiento del usuario y la sensibilidad de la acción solicitada. Un acceso desde el dispositivo habitual del usuario en su ubicación habitual para reproducir contenido requiere, según este modelo, un nivel de verificación distinto al que corresponde a un acceso desde un dispositivo nuevo en un país diferente para modificar el método de pago. Netflix, por ejemplo, implementó en 2022 un sistema de autenticación adaptativa que redujo en un 40% los accesos fraudulentos mediante account takeover, manteniendo una tasa de falsa activación —casos en que se exige verificación adicional a un usuario legítimo— inferior al 3% (Netflix Technology Blog, 2022).

Sin embargo, tres categorías de ataque limitan la eficacia de la autenticación incluso en sus formas más avanzadas. La primera es el secuestro de sesión (session hijacking), que no requiere las credenciales del usuario porque actúa sobre una sesión ya autenticada: el atacante intercepta el token de sesión —mediante un ataque man-in-the-middle, la explotación de vulnerabilidades en la implementación de TLS/HTTPS, o el sniffing en redes inseguras— y lo reutiliza desde su propia infraestructura, heredando los privilegios del usuario legítimo sin necesidad de conocer su contraseña ni su segundo factor de autenticación (Casey, 2011). La segunda es la ingeniería social aplicada a la autenticación: en el escenario del smishing analizado en el Capítulo 4, el atacante no necesita vulnerar el sistema de MFA técnicamente; basta con convencer al usuario de que introduzca su código OTP en una página fraudulenta porque, de lo contrario, su cuenta será bloqueada. Cialdini (2009) documenta que la combinación de urgencia temporal y autoridad de marca reconocida genera en el usuario un estado de cierre cognitivo bajo presión que inhibe la verificación crítica del mensaje recibido, haciendo que incluso usuarios con alta alfabetización digital caigan en este tipo de trampa. La tercera categoría es el compromiso del dispositivo final: si el dispositivo del usuario contiene un keylogger o un troyano de overlay —que superpone una interfaz falsa sobre la legítima en el momento del acceso—, la MFA más robusta puede ser capturada en el punto de entrada antes de que llegue al servidor (Kaspersky, 2025). El informe de amenazas móviles de Kaspersky del segundo trimestre de 2025 documentó 1.077 paquetes de instalación únicos de troyanos bancarios móviles dirigidos a usuarios europeos, un incremento del 23% respecto al mismo periodo de 2024.

La conclusión operativa de estas limitaciones no es que la autenticación reforzada sea inútil —los datos citados demuestran que es altamente eficaz cuando se implementa correctamente— sino que debe integrarse con controles de las capas superiores que detecten el abuso de sesiones ya autenticadas y el comportamiento anómalo post-login. La autenticación es condición necesaria pero no suficiente de la seguridad en plataformas de contenido.

6.2.2. Detección de anomalías y análisis de comportamiento

Los sistemas de detección de anomalías abordan la limitación fundamental de la autenticación: que esta solo verifica la identidad en el momento del acceso, pero no monitoriza si el comportamiento posterior es coherente con el del usuario legítimo. Un atacante que ha conseguido las credenciales de una cuenta y ha superado la MFA mediante ingeniería social tiene acceso ilimitado a los recursos de esa cuenta mientras el sistema no detecte que su comportamiento difiere del patrón histórico del usuario legítimo. Los sistemas de detección de anomalías son precisamente el mecanismo diseñado para cerrar esta ventana.

El análisis de comportamiento de usuarios y entidades (UEBA, User and Entity Behavior Analytics) es la tecnología de referencia en este espacio. Los sistemas UEBA construyen modelos de comportamiento habitual para cada usuario y entidad —perfiles de acceso, patrones de consumo, frecuencia de interacción, dispositivos utilizados, horarios— y generan alertas cuando el comportamiento observado se desvía estadísticamente de esos perfiles. Chandola et al. (2009) proporcionaron en su revisión seminal de técnicas de detección de anomalías el marco teórico que subyace a estos sistemas, distinguiendo entre anomalías puntuales (una sola transacción inusual), anomalías contextuales (una transacción normal en otro contexto pero inusual dado el historial del usuario) y anomalías colectivas (un patrón de transacciones que individualmente son normales pero que conjuntamente son indicativas de fraude). Esta última categoría —la anomalía colectiva— es particularmente relevante para el fraude en plataformas de contenido, donde los ataques de credential stuffing generan decenas de accesos desde IPs distintas que individualmente podrían interpretarse como accesos legítimos desde distintos dispositivos del usuario, pero que en conjunto revelan la firma de un ataque automatizado.

Las plataformas de mayor escala han desarrollado implementaciones específicas de UEBA orientadas a los tipos de fraude más prevalentes en su ecosistema. Spotify implementó a partir de 2021 un sistema de detección de fraude en streaming que analiza patrones de reproducción —duración media, frecuencia de cambio de pista, distribución geográfica de los oyentes, proporción entre reproducciones orgánicas e inducidas algorítmicamente— para identificar cuentas que inflan artificialmente los streams de ciertos artistas para maximizar los royalties que estos reciben. El sistema detectó y eliminó más de 35 millones de reproducciones fraudulentas en 2021, protegiendo la integridad del sistema de distribución de ingresos de la plataforma (Spotify, 2022). YouTube, por su parte, procesa más de 500 horas de vídeo subidas por minuto y aplica modelos de clasificación de tráfico que analizan en tiempo real la autenticidad de las visualizaciones, eliminando en 2023 más de 4.000 millones de reproducciones fraudulentas generadas por click farms y botnets, según su Informe de Transparencia (YouTube, 2023).

La detección de actividades coordinadas entre múltiples cuentas añade una dimensión de análisis de grafos a los modelos de anomalía individual. En lugar de analizar el comportamiento de cada cuenta de forma aislada, los sistemas de análisis de grafos construyen representaciones de la red de interacciones entre cuentas y detectan patrones de coordinación que son indicativos de comportamiento fraudulento: grupos de cuentas que se siguen mutuamente en patrones regulares, que publican contenido similar en ventanas temporales estrechas, o que interactúan con el mismo conjunto reducido de cuentas objetivo. Akoglu et al. (2015) formalizaron las técnicas de análisis de grafos aplicadas a la detección de fraude en redes, demostrando que los grafos de comportamiento de redes de cuentas fraudulentas presentan propiedades topológicas —densidad de conexiones, distribución de grados, coeficiente de clustering— estadísticamente distintas de las de las redes orgánicas de usuarios legítimos. Facebook utilizó técnicas de análisis de grafos para identificar y eliminar más de 2.800 millones de cuentas falsas en 2018, una operación que su equipo de seguridad documentó como la mayor desarticulación de infraestructura de fraude coordinado en la historia de la plataforma (Facebook, 2018).

A pesar de su eficacia documentada, los sistemas de detección de anomalías presentan tres limitaciones estructurales que deben gestionarse activamente. La primera es la generación de falsos positivos: cualquier modelo que detecta desviaciones del

comportamiento habitual bloqueará inevitablemente a usuarios legítimos que exhiben comportamientos inusuales por razones inocentes —viajes internacionales, cambios de dispositivo, picos de actividad en momentos de ocio—. Phua et al. (2010) analizaron el trade-off entre tasa de detección y tasa de falsos positivos en sistemas de detección de fraude y concluyeron que los modelos operativos deben calibrarse con umbrales de alerta que maximicen el daño evitado neto, ponderando el coste del fraude no detectado contra el coste de la fricción impuesta a usuarios legítimos. Para plataformas de contenido con bases de usuarios de decenas de millones, incluso una tasa de falsos positivos del 0,1% genera cientos de miles de incidencias diarias de usuarios legítimos bloqueados, con el consiguiente impacto en la experiencia de usuario y en el coste operativo de los equipos de soporte. La segunda limitación es la dependencia de la calidad de los datos de entrenamiento: los modelos de detección son tan buenos como los datos históricos con los que se entrenan, y si esos datos contienen sesgos —por ejemplo, si el historial de fraude detectado está subrepresentado en ciertos segmentos demográficos o geográficos—, el modelo replicará esos sesgos en sus predicciones, generando tanto infradetección en los segmentos no representados como sobredetección en los más representados. La tercera limitación es la adaptación progresiva de los atacantes: las organizaciones criminales analizan los patrones de bloqueo que generan sus ataques y modifican sus técnicas para emular más fielmente el comportamiento legítimo, en un ciclo de coevolución atacante-defensor que el modelo conceptual de la Sección 9.2.3.4 de este trabajo formaliza como dinámica adaptativa (Baxter & Sommerville, 2011).

6.2.3. Sistemas de protección perimetral y de aplicaciones

Los Web Application Firewalls (WAF) y los sistemas de protección perimetral constituyen la tercera capa de defensa, orientada a filtrar el tráfico malicioso antes de que alcance las capas de aplicación y datos de la plataforma. Su función principal es inspeccionar las solicitudes HTTP/HTTPS entrantes en busca de patrones que coincidan con firmas de ataques conocidos —inyecciones SQL, cross-site scripting (XSS), inclusión de archivos remotos (RFI), ataques de Directory Traversal— y bloquear aquellas que suponen un riesgo para la integridad de la aplicación. En el contexto del fraude en plataformas de contenido, los WAF son especialmente relevantes para mitigar dos tipos de ataque: el scraping masivo de contenido —que puede constituir tanto una vulneración de los derechos de propiedad intelectual de la plataforma como un mecanismo de captura de datos de usuarios para ataques posteriores— y el abuso de APIs, que permite a los atacantes automatizar acciones que generarían comportamientos fraudulentos a escala.

El abuso de APIs representa un vector de ataque de creciente relevancia a medida que las plataformas de contenido exponen más funcionalidades a través de interfaces programáticas. Las APIs públicas o semipúblicas que permiten la búsqueda de contenido, la gestión de playlists o la recuperación de datos de usuario pueden ser utilizadas de forma automatizada para ejecutar ataques de credential stuffing a velocidades muy superiores a las que sería posible mediante la interfaz de usuario web, para extraer bases de datos de usuarios mediante scraping sistemático, o para generar tráfico artificial sobre contenidos específicos. La respuesta técnica incluye la implementación de rate limiting —límites al número de solicitudes por unidad de tiempo por cliente o token de acceso—, la validación estricta de los tokens de autenticación de la API, el análisis del comportamiento de las llamadas a la API para detectar patrones automatizados, y el bloqueo geográfico o de rangos de IP cuando el tráfico es inconsistente con el perfil esperado del cliente legítimo de la API.

Una limitación estructural de los sistemas de protección perimetral es que operan principalmente a nivel de tráfico y firma de ataque, no a nivel de intención o de semántica del comportamiento. Un sistema WAF puede detectar y bloquear una solicitud de SQL injection porque coincide con una firma conocida, pero no puede distinguir entre un acceso legítimo a la API de búsqueda y un ataque de scraping sistemático si ambos generan solicitudes sintácticamente válidas. Esta limitación explica por qué los sistemas perimetrales deben complementarse inevitablemente con los sistemas de análisis de comportamiento descritos en la sección anterior: la combinación de filtrado por firma en el perímetro con análisis de anomalías en el comportamiento de la aplicación proporciona una cobertura que ninguno de los dos mecanismos ofrece de forma aislada. ENISA, en su análisis de las arquitecturas de seguridad de plataformas de contenido, recomienda explícitamente la integración de WAF con sistemas UEBA como estándar mínimo de protección para plataformas con más de un millón de usuarios activos (ENISA, 2023a).

Los sistemas de CAPTCHA y las herramientas de verificación de interacción humana representan un caso particular en esta capa, específicamente orientado a impedir el acceso automatizado a funcionalidades de la plataforma que requieren comportamiento humano. Los CAPTCHA estáticos de primera generación —reconocimiento de texto distorsionado— fueron superados por soluciones de resolución automatizada basadas en modelos de visión computacional en la primera mitad de la década de 2010. Los sistemas de verificación de tercera generación actualmente desplegados —como reCAPTCHA v3 de Google y los sistemas de análisis de fingerprint de navegador— evalúan la probabilidad de que el usuario sea humano de forma transparente, sin interrumpir el flujo de interacción, analizando cientos de variables de comportamiento micro —movimiento del ratón, velocidad de escritura, patrones de desplazamiento, características del entorno de ejecución del navegador— para construir una puntuación de riesgo en tiempo real. Su eficacia contra los sistemas de resolución automatizada más sofisticados sigue siendo objeto de debate académico activo, con estudios recientes que documentan tasas de evasión superiores al 85% para los modelos de resolución más avanzados (Searles et al., 2023), lo que sugiere que su valor defensivo residual se encuentra más en el incremento del coste y la complejidad del ataque que en su capacidad de bloquearlo de forma absoluta.

6.2.4. Sistemas antifraude basados en inteligencia artificial

La inteligencia artificial ha transformado el espacio de la detección de fraude en plataformas digitales en los últimos cinco años, desplazando los sistemas basados en reglas estáticas —que requieren la definición explícita de cada patrón de fraude por parte de analistas humanos— hacia modelos de aprendizaje automático capaces de identificar patrones complejos y emergentes de forma autónoma. Esta transformación ha incrementado sustancialmente la capacidad de detección de fraudes conocidos y, en menor medida, de fraudes nuevos cuya firma no ha sido previamente catalogada.

Los modelos supervisados son los más ampliamente desplegados en producción. Entrenados sobre conjuntos de datos etiquetados que distinguen transacciones o comportamientos legítimos de fraudulentos, aprenden las características estadísticas que discriminan entre ambas clases y las aplican a nuevas instancias en tiempo real. Los algoritmos más utilizados en este contexto son los árboles de decisión con boosting (XGBoost, LightGBM) para clasificación de transacciones de pago, las redes neuronales profundas para clasificación de comportamiento en plataformas con alta dimensionalidad de señales, y los modelos de secuencia (LSTM, Transformers) para la detección de

patrones temporales en el comportamiento del usuario. Phua et al. (2010) establecieron en su revisión comprehensiva de técnicas de detección de fraude basadas en minería de datos los criterios de evaluación —precisión, recall, área bajo la curva ROC— que siguen siendo el estándar de evaluación de estos modelos. Meta informó en 2023 que sus modelos supervisados de detección de comportamiento inauténtico coordinado identificaban más del 90% de las campañas de desinformación y fraude en la plataforma antes de que alcanzaran escala significativa, con una tasa de falsos positivos inferior al 0,01% del contenido legítimo (Meta, 2023).

Los modelos no supervisados abordan el problema fundamental de los supervisados: la dependencia de datos etiquetados que, por definición, solo pueden reflejar los patrones de fraude ya conocidos en el momento del etiquetado. Técnicas como el clustering de k-medias, los autoencoders y los modelos de mezcla gaussiana identifican agrupaciones y anomalías en los datos sin necesidad de etiquetado previo, permitiendo la detección de nuevos tipos de fraude cuya firma aún no ha sido incorporada a los modelos supervisados. Su principal limitación es la dificultad de interpretación: mientras que un modelo supervisado puede explicar por qué clasifica una transacción como fraudulenta —por ejemplo, porque la IP de origen pertenece a un rango conocido de fraude, o porque el importe supera en tres desviaciones estándar la media histórica del usuario—, un modelo no supervisado identifica que algo es anómalo pero no proporciona una explicación semántica que permita al analista humano validar la alerta (Chandola et al., 2009).

El análisis de grafos con aprendizaje automático representa la contribución más original de la IA al espacio de la detección de fraude en plataformas de contenido, porque aborda un tipo de patrón que los modelos clásicos —que analizan instancias individuales— no pueden detectar: las relaciones entre cuentas que revelan coordinación fraudulenta. Akoglu et al. (2015) demostraron que los grafos de interacción de redes de cuentas fraudulentas presentan propiedades topológicas identificables mediante técnicas de graph embedding y clasificación de grafos. Aplicaciones concretas en plataformas de contenido incluyen la detección de redes de seguidores falsos en redes sociales —identificando grupos de cuentas que se siguen mutuamente en patrones regulares inconsistentes con la formación orgánica de redes sociales—, la identificación de coordinadas de cuentas que inflan métricas de contenido específico —vídeos, canciones, artículos— de forma sincronizada, y la detección de infraestructuras de fraude publicitario que utilizan redes de dispositivos comprometidos para generar clics fraudulentos en anuncios.

Los desafíos más relevantes de los sistemas de IA aplicados al fraude en plataformas de contenido son de naturaleza tanto técnica como regulatoria. Técnicamente, la opacidad algorítmica de los modelos de aprendizaje profundo genera dificultades de auditoría y explicación: cuando un modelo bloquea una cuenta o rechaza una transacción, ni el usuario afectado ni el regulador pueden entender fácilmente qué variables determinaron esa decisión, lo que puede constituir una violación del derecho de los interesados a una explicación de las decisiones automatizadas reconocido en el artículo 22 del RGPD (Unión Europea, 2016). El AI Act europeo (Unión Europea, 2024a) clasifica los sistemas de IA que toman decisiones que afectan significativamente a los derechos de los ciudadanos como de alto riesgo, imponiéndoles requisitos de transparencia, robustez y supervisión humana que muchos sistemas de detección de fraude actuales no cumplen en su totalidad. Desde la perspectiva del riesgo de sesgo, los modelos entrenados con datos históricos de fraude pueden replicar y amplificar patrones de sobredetección en determinados grupos demográficos si los datos de entrenamiento contienen sesgos en la distribución de los casos etiquetados como fraudulentos (ENISA, 2023b).

6.3. Investigación digital y análisis forense

6.3.1. Obtención y preservación de evidencia digital

Cuando los controles preventivos y de detección no han sido suficientes para evitar un incidente de fraude, la fase de investigación digital determina en qué medida es posible reconstruir lo ocurrido, identificar a los responsables, limitar el daño continuado y sustentar las acciones legales oportunas. La investigación de fraude en plataformas de contenido digital presenta particularidades que la distinguen de la investigación forense en entornos corporativos tradicionales: la evidencia está distribuida entre múltiples actores —la plataforma, el operador telco, la entidad financiera, los registradores de dominios—, los sistemas investigados operan a escala masiva con logs que generan varios terabytes de datos diarios, y los atacantes utilizan sistemáticamente técnicas de anonimización y volatilidad de infraestructura que hacen que la evidencia más valiosa tenga una vida útil muy corta.

Casey (2011) estableció en su referencia seminal sobre evidencia digital los principios fundamentales que deben regir la obtención y preservación de evidencia en cualquier investigación digital: legalidad, proporcionalidad, integridad y reproducibilidad. La legalidad exige que la recopilación de evidencia se realice conforme a la normativa aplicable, lo que en el contexto de plataformas de contenido implica navegar la tensión entre el imperativo de preservar la evidencia y los requisitos del RGPD sobre minimización de datos y limitación de la finalidad del tratamiento (Unión Europea, 2016). La proporcionalidad establece que la recopilación debe limitarse a los datos estrictamente necesarios para la investigación, lo que requiere un análisis previo de qué tipos de evidencia son relevantes para el tipo específico de fraude investigado. La integridad exige que la evidencia obtenida no sea alterada en el proceso de recopilación o preservación, lo que en la práctica se garantiza mediante el cálculo de funciones hash criptográficas (SHA-256 o SHA-3) sobre los archivos de log o imágenes forenses en el momento de su obtención, permitiendo demostrar en cualquier momento posterior que la evidencia no ha sido modificada. La reproducibilidad exige que el proceso de obtención de evidencia sea documentado con suficiente detalle como para que otro investigador pueda repetirlo y llegar a los mismos resultados, lo que constituye la cadena de custodia.

Los tipos de evidencia más relevantes en la investigación de fraude en plataformas de contenido incluyen cuatro categorías. Los registros de acceso y actividad (logs) constituyen la fuente primaria de la investigación: contienen los registros de las direcciones IP de origen, los timestamps de cada acción, los identificadores de sesión, los dispositivos utilizados (inferidos del User-Agent del navegador) y la secuencia de acciones realizadas dentro de la sesión. Su principal limitación forense es la duración de la retención: la mayoría de las plataformas retienen los logs de acceso entre treinta y noventa días, lo que significa que investigaciones que comienzan tarde pueden encontrarse con que la evidencia más relevante ya ha sido eliminada conforme a las políticas de retención. Los metadatos de transacción incluyen los registros de cambios de cuenta, adquisiciones de suscripciones, cambios de método de pago y transferencias, con toda la información temporal, de origen y de contexto de cada operación. La información de dispositivo —fingerprint del navegador, identificadores de hardware accesibles desde el navegador, patrones de resolución de pantalla y fuentes instaladas— permite en muchos casos atribuir múltiples cuentas a un mismo dispositivo físico, identificando redes de cuentas operadas desde una misma infraestructura. Los datos de geolocalización,

finalmente, permiten correlacionar accesos desde ubicaciones inconsistentes con el perfil del usuario legítimo y detectar el uso de VPN o proxies cuando la IP de origen pertenece a un rango de proveedores de anonimización conocidos.

La preservación de evidencia en investigaciones de fraude en plataformas de contenido digital enfrenta un desafío que Casey (2011) ya identificó como crítico: la volatilidad de la infraestructura criminal. Los atacantes modernos utilizan infraestructuras efímeras —dominios de phishing registrados días antes del ataque y abandonados horas después, servidores de command-and-control con ciclos de vida de horas, cuentas de plataformas de comunicación eliminadas en cuanto se completa el ciclo de fraude— precisamente para maximizar la dificultad de la preservación forense. Esto exige que los procedimientos de preservación se activen en tiempo casi real una vez detectado el incidente, y que existan acuerdos de cooperación preestablecidos entre las plataformas y los cuerpos de seguridad que permitan la solicitud y ejecución de órdenes de preservación de evidencia en plazos compatibles con la volatilidad de los datos.

6.3.2. Análisis forense y atribución

El objetivo del análisis forense en casos de fraude en plataformas de contenido es triple: reconstruir la secuencia de eventos con la mayor precisión cronológica posible, identificar los vectores de ataque y las vulnerabilidades explotadas, y atribuir la autoría con la solidez técnica suficiente para sustentar una acción legal. En la práctica, los tres objetivos son interdependientes: la reconstrucción cronológica es el insumo del análisis de vectores, y ambos son la base de la atribución.

La correlación de logs entre sistemas es la técnica fundamental de la reconstrucción cronológica. En investigaciones de fraude en plataformas de contenido, los eventos relevantes están distribuidos en logs de sistemas distintos —el servidor web de la plataforma, el sistema de gestión de sesiones, el sistema de pagos, los logs del operador telco si el vector fue smishing, los registros de la entidad financiera si hubo fraude de pago— que utilizan relojes con distintos niveles de precisión y sincronización. La normalización temporal de todos los logs —usando UTC como referencia y ajustando las marcas temporales de cada sistema a esa referencia— es el primer paso del análisis, y su importancia es crítica: una diferencia de un segundo en la correlación temporal puede cambiar completamente la secuencia causal de los eventos. El software forense especializado como Plaso/log2timeline (Gudjonsson, 2012) automatiza este proceso para los formatos de log más comunes, pero las plataformas de contenido frecuentemente utilizan formatos propietarios que requieren parsers específicos.

El análisis de redes en el contexto forense difiere del análisis de redes aplicado a la detección de fraude descrito en la Sección 6.2.2: mientras que en la detección el objetivo es identificar patrones anómalos en tiempo real sobre grandes volúmenes de datos, en el análisis forense el objetivo es reconstruir la infraestructura específica utilizada en un ataque concreto, identificando los nodos —IPs, dominios, cuentas— que formaron parte de la operación fraudulenta y las relaciones entre ellos. La resolución pasiva de DNS —consultando los registros históricos de los servidores de nombres para reconstruir qué IPs resolvió un dominio en un momento específico— y el análisis de los registros WHOIS y de los certificados TLS —que pueden revelar relaciones entre dominios registrados por el mismo operador— son técnicas forenses estándar en la atribución de infraestructura criminal en casos de fraude en plataformas digitales (Europol, 2023b).

La inteligencia de fuentes abiertas (OSINT) completa el análisis forense proporcionando contexto sobre la infraestructura y los actores identificados. Servicios como Shodan —que indexa dispositivos conectados a internet y sus servicios expuestos—, VirusTotal —que agrega análisis de archivos y URLs de múltiples motores antivirus— y URLscan.io —que captura screenshots y metadatos de páginas web en el momento de su análisis— permiten correlacionar la infraestructura técnica de un ataque con campañas anteriores o con actores conocidos. En investigaciones de fraude en plataformas de contenido, el análisis OSINT de los dominios fraudulentos identificados puede revelar que fueron registrados por el mismo correo electrónico o mediante el mismo registrador que dominios utilizados en ataques anteriores, proporcionando un hilo de atribución que trasciende el incidente individual.

El reto más difícil del análisis forense en fraude de plataformas de contenido sigue siendo la atribución definitiva a personas físicas. Las técnicas de anonimización —VPN comerciales, nodos Tor, proxies residenciales que utilizan dispositivos de consumidores domésticos como puntos de salida— hacen que la IP de origen visible en los logs de la plataforma rara vez corresponda a la ubicación real del atacante. La atribución efectiva requiere frecuentemente la cooperación de múltiples actores: el proveedor de VPN que tiene los registros de qué IP real se conectó al servidor VPN en el momento del ataque, el operador telco que tiene los registros de la conexión de esa IP real, y en última instancia la autoridad policial del país de origen del atacante. Esta cadena de cooperación, que implica solicitudes de asistencia judicial internacional, puede durar meses o años, tiempo durante el cual los datos relevantes pueden haber expirado en los sistemas de cada intermediario. La Directiva NIS2 (Unión Europea, 2022) introduce obligaciones de conservación de logs relevantes para la seguridad de las redes que reducen parcialmente este problema para los operadores europeos, pero no resuelve la parte transnacional de la cadena de atribución.

6.4. Colaboración público-privada en la respuesta al fraude

Los controles técnicos y forenses descritos en las secciones anteriores operan, en su mayor parte, dentro del perímetro de una sola organización. Su eficacia frente a los tipos de fraude más sofisticados —los que combinan vectores de telecomunicaciones, plataforma digital y sistema financiero, como se ilustra en el Capítulo 8 — está fundamentalmente limitada por la ausencia de mecanismos que permitan correlacionar señales entre organizaciones en tiempo real y activar respuestas coordinadas. La colaboración público-privada es el mecanismo que transforma controles aislados en un sistema de defensa cohesionado.

Los equipos de respuesta a incidentes de ciberseguridad (CSIRT/CERT) constituyen la infraestructura de coordinación técnica más consolidada en este espacio. En el ecosistema europeo, los CSIRTs nacionales operan en el marco de la red CSIRTs establecida por la Directiva NIS2 (Unión Europea, 2022), que obliga a los Estados miembros a designar uno o más CSIRTs con mandato sobre diferentes categorías de operadores de servicios esenciales y proveedores de servicios digitales. INCIBE-CERT es el CSIRT de referencia para ciudadanos y entidades privadas en España, mientras que el CCN-CERT tiene competencia sobre las administraciones públicas y los operadores de infraestructuras críticas. Sin embargo, el alcance del mandato de los CSIRTs sobre el fraude en plataformas de contenido —que no es un incidente de seguridad en el sentido técnico de NIS2, sino un fenómeno que combina fraude financiero, uso malicioso de infraestructuras

de telecomunicaciones y abuso de plataformas digitales— no está claramente delimitado, lo que genera una zona gris en la que la coordinación es más difícil.

El intercambio de inteligencia de amenazas entre los actores del ecosistema es la forma más operativamente relevante de colaboración público-privada en el contexto del fraude en plataformas de contenido. Este intercambio puede articularse mediante dos mecanismos complementarios. El primero es el intercambio reactivo: cuando un actor detecta un incidente de fraude, notifica a los demás actores potencialmente afectados para que puedan bloquear los indicadores de compromiso identificados (IPs, dominios, hashes de malware, números de teléfono fraudulentos) antes de que el ataque les alcance. El segundo es el intercambio proactivo: la participación en plataformas de threat intelligence sharing —como MISP o las ISACs sectoriales— que agregan indicadores de múltiples actores y los diseminan de forma normalizada, permitiendo que los sistemas de detección de cada organización se beneficien de la inteligencia colectiva del ecosistema (Europol, 2023a). El GSMA Fraud Intelligence Sharing Service (FISS) es el mecanismo de intercambio proactivo de referencia para el sector de telecomunicaciones, con datos del GSMA Fraud Management Framework (2025) que muestran que los operadores participantes en el FISS detectan las campañas de fraude telco en promedio 6,4 horas antes que los no participantes.

Los procedimientos conjuntos de investigación entre plataformas y fuerzas de seguridad son el mecanismo de colaboración más complejo de implementar, porque implican la superación de barreras legales, organizativas y de incentivos que son más difíciles de resolver que las meramente técnicas. Europol articula estos procedimientos a nivel supranacional mediante la Joint Cybercrime Action Taskforce (J-CAT), que en 2023 coordinó 21 operaciones de gran escala contra infraestructuras de fraude digital, resultando en la detención de 86 personas y el desmantelamiento de infraestructuras responsables de pérdidas estimadas en más de 400 millones de euros (Europol, 2023b). A nivel doméstico, la Brigada de Investigación Tecnológica de la Policía Nacional española y la UCO de la Guardia Civil tienen competencias complementarias en la investigación de fraude digital, aunque la coordinación entre ambas unidades y con los actores privados del ecosistema de contenido sigue siendo una asignatura pendiente en la arquitectura institucional española.

La colaboración público-privada en materia de fraude enfrenta cuatro desafíos estructurales que ningún actor puede resolver de forma unilateral. El primero son las diferencias regulatorias entre jurisdicciones: los actores del ecosistema operan bajo marcos legales distintos que determinan qué datos pueden compartirse, con quién y bajo qué condiciones, generando asimetrías que los atacantes explotan deliberadamente localizando su infraestructura en jurisdicciones con menor cooperación judicial. El segundo es el conflicto entre privacidad y seguridad: los datos necesarios para una coordinación efectiva en la prevención del fraude —patrones de comportamiento de usuarios, señales de tráfico de red, registros de transacciones— son también datos personales protegidos por el RGPD, y su intercambio requiere habilitaciones jurídicas específicas que no siempre están disponibles o claramente interpretadas. El tercero es la asimetría de incentivos: las plataformas y operadores tienen incentivos para no revelar públicamente los incidentes de fraude que afectan a sus usuarios, porque dicha revelación puede generar daño reputacional y responsabilidad legal, mientras que los reguladores y fuerzas de seguridad tienen incentivos para exigir transparencia. El cuarto es la asimetría de capacidades: las organizaciones con mayores capacidades técnicas —las grandes plataformas con equipos de seguridad de cientos de personas— generan la mayor parte

de la inteligencia sobre fraude, pero no necesariamente tienen los incentivos para compartirla con actores más pequeños que son sus competidores en el mercado. La superación de estos cuatro desafíos requiere un diseño institucional deliberado que alinee los incentivos de todos los actores hacia la cooperación, que es precisamente el objetivo de la Plataforma Nacional de Inteligencia Antifraude propuesta en la Recomendación R1 de la Sección 7.1.

6.5. Síntesis: hacia una arquitectura de defensa integrada

Los cuatro controles analizados en esta sección —autenticación, detección de anomalías, protección perimetral e investigación forense— son complementarios y mutuamente dependientes. La autenticación reduce el número de sesiones comprometidas que llegan a la capa de detección de anomalías; la detección de anomalías identifica los comportamientos fraudulentos que superan la capa de autenticación; la protección perimetral filtra el tráfico automatizado que intenta eludir ambas capas; y la investigación forense cierra el ciclo proporcionando la retroalimentación necesaria para mejorar los controles preventivos. Ninguno de estos mecanismos, operando de forma aislada, puede contener el fraude adaptativo descrito en el Capítulo 4; su eficacia emerge de su integración en una arquitectura coherente.

Esa arquitectura, sin embargo, no puede limitarse al perímetro de una sola organización. La convergencia entre sectores que caracteriza el fraude moderno —documentada a lo largo de este trabajo— exige extender la arquitectura de defensa a través del ecosistema, conectando los controles de las plataformas con los de los operadores telco y las entidades financieras mediante los mecanismos de colaboración público-privada descritos en la Sección 6.4. El GSMA Fraud Management Framework (2025) propone un modelo de madurez para esta arquitectura de defensa interorganizacional, en el que los operadores progresan desde un nivel inicial de defensa perimetral aislada hasta un nivel avanzado de inteligencia compartida y respuesta coordinada en tiempo real. Según los datos del GSMA Global Fraud Loss Survey (FASG#34, 2026), los operadores en el nivel más alto de madurez del modelo registran pérdidas por fraude un 62% inferiores a las de los operadores en el nivel más bajo, lo que cuantifica el valor económico de la integración. España se encuentra, en la evaluación del GSMA para el ecosistema europeo, en una posición intermedia de este espectro de madurez, con capacidades técnicas bien desarrolladas en autenticación y detección de anomalías pero con importantes déficits en los mecanismos de inteligencia compartida intersectorial y de coordinación operativa. Cerrar esos déficits es el objetivo estratégico de las Recomendaciones R1 a R4 del Capítulo 7.

7. Recomendaciones estratégicas prioritarias y rol de INCIBE

Este capítulo presenta las recomendaciones estratégicas prioritarias para la mitigación del fraude digital, organizadas en cuatro ejes: gobernanza e inteligencia compartida, regulación y responsabilidad de plataformas, capacidades técnicas, y educación y resiliencia del usuario. Se incluye además un marco de métricas e indicadores de seguimiento y el papel de INCIBE en el ecosistema antifraude.

Las recomendaciones que se formulan a continuación son propuestas de acción específicas, fundadas en el análisis desarrollado a lo largo de este trabajo, dirigidas a actores concretos del ecosistema español y calibradas en función de los déficits que el análisis ha identificado. Cada recomendación se articula en torno a tres elementos: el problema específico que resuelve, con referencia al capítulo del trabajo donde se documenta; la propuesta de actuación con detalle operativo suficiente para que sea implementable; y al menos un ejemplo o referencia internacional que muestra que la propuesta es factible en condiciones análogas a las del contexto español.

7.1. Gobernanza e inteligencia compartida

R1 — Crear la Plataforma Nacional de Inteligencia Antifraude (PNIA)

El principal déficit estructural del ecosistema español es la ausencia de una infraestructura de intercambio de inteligencia sobre fraude que opere entre sectores en tiempo real. El resultado, como se documenta en la Sección 4.7 y se ilustra en el Capítulo 8, es que los ataques más sofisticados —los que combinan vectores telco, plataforma y financiero— son sistemáticamente los más difíciles de detectar y los que generan mayor daño, precisamente porque explotan las discontinuidades entre los sistemas de información de actores que no se comunican (GSMA, 2025; Europol, 2023b).

La PNIA resuelve este problema proporcionando una infraestructura técnica neutral, operada por INCIBE, que actúa como hub de intercambio de indicadores de fraude bajo un modelo federado. El estándar técnico de referencia es MISP, que Europol ha adoptado como plataforma central de su programa de intercambio de inteligencia sobre cibercrimen (Europol, 2023a) y que utilizan más de ochenta CSIRTs europeos. La adopción de MISP como base de la PNIA no requiere desarrollar tecnología nueva: requiere acordar la taxonomía de indicadores de fraude, definir los flujos de ingesta desde los actores participantes y establecer el protocolo de acceso diferenciado. La taxonomía debería alinearse con la del GSMA Fraud Management Framework v22.0 (GSMA, 2025) y la del CFCA Global Fraud Loss Survey (CFCA, 2023), que son los estándares de referencia internacionales en el sector de telecomunicaciones.

El modelo de referencia más cercano en términos de alcance y contexto es el ISAC del sector financiero europeo articulado por el Banco Central Europeo bajo el programa TIBER-EU. En cinco años, este ISAC logró que el 78% de las entidades financieras significativas de la zona euro participaran activamente en el intercambio de indicadores de fraude (BCE, 2022). La clave del éxito fue la resolución temprana de la cuestión jurídica, la participación graduada y una gobernanza neutral. Estos tres elementos deben estar presentes en el diseño de la PNIA desde el primer momento.

R2 — Establecer obligaciones de reporte estandarizado de incidentes de fraude

El intercambio voluntario de inteligencia tiene un límite estructural conocido: los actores comparten la información que les conviene compartir y retienen la que consideran sensible. La solución es la obligatoriedad del reporte, proporcional, estandarizada y operativamente manejable. La propuesta es ampliar el alcance de las obligaciones de notificación de incidentes ya existentes bajo la Directiva NIS2 (Unión Europea, 2022) para incluir los incidentes de fraude digital con impacto significativo sobre usuarios finales. El umbral de notificación podría definirse como cualquier incidente que afecte a más de 1.000 usuarios identificados, genere pérdidas directas superiores a 100.000 euros, o utilice infraestructura técnica activa en el momento de la notificación. La taxonomía de reporte se alinearía con la del GSMA Fraud Management Framework (GSMA, 2025) y con los criterios de reporte de ENISA bajo NIS2 (ENISA, 2023c).

La experiencia del Mandatory Fraud Reporting Scheme del Reino Unido, introducido por la Online Safety Act 2023, es ilustrativa. En su primer año de funcionamiento, el esquema generó un incremento del 240% en el volumen de inteligencia disponible sobre fraude en plataformas digitales, lo que permitió identificar cuatro redes criminales no detectadas previamente (Home Office, 2024). El principal problema operativo fue la sobrecarga inicial de los sistemas de reporte, que debe anticiparse priorizando la calidad sobre la cantidad de los reportes y estableciendo desde el primer momento los mecanismos de agregación y filtrado.

R3 — Integrar el fraude digital como riesgo estratégico en la política nacional de ciberseguridad

La Estrategia Nacional de Ciberseguridad vigente en España, aprobada en 2019 y actualizada en 2022 (Departamento de Seguridad Nacional, 2022), aborda el fraude digital de forma tangencial, principalmente como manifestación del cibercrimen dirigido a infraestructuras críticas. No lo trata como un riesgo sistémico de primer orden para la economía digital, que es lo que los datos de los últimos tres años evidencian: 106.800 infracciones penales por estafas informáticas en el primer trimestre de 2025 y pérdidas superiores a 350 millones de euros anuales (Ministerio del Interior, 2026; Visa España, 2025). La integración propuesta requiere incorporar este vector de riesgo en el próximo ciclo de actualización estratégica, con líneas de acción específicas, indicadores de seguimiento y asignación presupuestaria, en coherencia con los marcos NIS2, AI Act y DSA que también abordan el fraude digital de forma transversal (Unión Europea, 2022, 2024a, 2024b).

7.2. Regulación y responsabilidad de plataformas

R4 — Desarrollar el marco de responsabilidad de plataformas en fraude distribuido bajo el DSA

El Reglamento de Servicios Digitales (DSA), en vigor desde febrero de 2024 para las plataformas de gran tamaño (Unión Europea, 2022b), introduce obligaciones de evaluación de riesgos sistémicos (artículo 34) y adopción de medidas de mitigación razonables (artículo 35). Sin embargo, la aplicación de estos artículos al fraude distribuido en plataformas de contenido está siendo objeto de debate interpretativo. La Comisión

Europea designó a Meta, Google, TikTok y Amazon como Very Large Online Platforms (VLOPs) y les exigió evaluaciones de riesgo sistémico específicas. Las evaluaciones publicadas por Meta y Google en 2024 revelan metodologías de identificación de riesgo de fraude que, aunque diseñadas para sus especificidades, ofrecen un punto de partida para definir estándares adaptables al conjunto del mercado (Comisión Europea, 2024). La propuesta es que la CNMC, como autoridad coordinadora del DSA en España, elabore en el plazo de dieciocho meses una guía de aplicación del artículo 34 al fraude en plataformas de contenido, clarificando los tipos de fraude que constituyen riesgo sistémico, los indicadores de detección mínimos por tipología de plataforma y los plazos de notificación ante incidentes de fraude que afecten a usuarios en España.

R5 — Desarrollar capacidades nacionales de detección de deepfakes aplicados al fraude

La aceleración del fraude basado en contenido sintético —documentada en la Sección 4.3 con el dato de duplicación del volumen de deepfakes aproximadamente cada seis meses (ISMS Forum, 2026; Keepnet Labs, 2026)— plantea un desafío que no puede resolverse únicamente con regulación. La detección de deepfakes requiere capacidades técnicas en visión computacional y análisis de audio que están en la frontera del estado del arte.

El caso de referencia es el programa de detección de deepfakes que DARPA financió entre 2019 y 2023 bajo el proyecto Media Forensics (MediFor), que desarrolló herramientas de detección de manipulación de contenido audiovisual y las puso a disposición de plataformas digitales y agencias gubernamentales (DARPA, 2023). La principal limitación fue que las herramientas desarrolladas envejecen rápidamente: los modelos de generación de deepfakes evolucionan más rápido que los de detección. Esto implica que la inversión debe ser un programa continuo de I+D, no un proyecto puntual. La propuesta para España es un consorcio de investigación en detección de contenido sintético, coordinado por INCIBE y financiado con fondos del programa Horizonte Europa y del PERTE Chip, que integre universidades con grupos de investigación en inteligencia artificial, operadores telco y empresas del sector de verificación de contenido. Las herramientas resultantes deberían licenciarse en abierto, alineándose con el mandato del artículo 50 del AI Act sobre transparencia de sistemas de IA generativa (Unión Europea, 2024a).

R6 — Establecer requisitos mínimos de autenticación en plataformas de contenido de alto riesgo

La autenticación reforzada del cliente (SCA) es obligatoria bajo PSD2 (Unión Europea, 2015) para las transacciones de pago, con niveles proporcionales al importe y al riesgo de la operación. Este criterio deja sin cubrir las plataformas de contenido digital, que gestionan datos personales de alto valor y métodos de pago vinculados sin requisitos equivalentes. La propuesta extiende el principio de autenticación proporcional al riesgo a las plataformas de contenido que superen umbrales de usuarios activos, estableciendo como mínimo: MFA obligatoria para modificaciones de cuenta críticas (cambios de contraseña, método de pago y correo electrónico), autenticación adaptativa para accesos desde nuevos dispositivos o ubicaciones, y notificación proactiva de actividades inusuales.

El argumento habitual de las plataformas contra la autenticación reforzada —deterioro de la experiencia de usuario— está siendo revisado a la luz de datos recientes. Google publicó en 2023 que la activación de MFA redujo el riesgo de account takeover en un

99,9% con un incremento de la tasa de abandono en el flujo de acceso inferior al 2% cuando se implementa con autenticación biométrica (Google, 2023). Spotify, que introdujo verificación adicional para cambios de método de pago en 2022, reportó una reducción del 68% en los fraudes de cambio de tarjeta con impacto mínimo en la satisfacción del usuario (Spotify, 2022). Estos datos sugieren que el trade-off entre seguridad y conversión es menos severo de lo que la industria asume, especialmente con diseños ergonómicos y proporcionales al nivel de riesgo de cada acción.

7.3. Capacidades técnicas de detección y prevención

R7 — Protocolo de detección colaborativa de cuentas mulas entre operadores telco y entidades financieras

La cuenta mula es el cuello de botella del ecosistema del fraude: sin ella, la monetización es imposible o extremadamente costosa. La paradoja es que su detección sigue siendo uno de los elementos más débiles del sistema, porque la señal definitiva de que una cuenta es una mula solo es visible cuando se correlacionan datos de diferentes sectores. Los patrones que permiten identificarla —actividad de SMS anómala en el operador telco, transferencias en importes justo por debajo del umbral de verificación automática en el banco, accesos desde múltiples IPs en la plataforma de streaming— son invisibles para cada actor de forma aislada pero inequívocos cuando se correlacionan. Según los modelos de detección del BioCatch Financial Consortium, la confluencia de estas tres señales identifica una cuenta mula activa con una probabilidad superior al 95% (BioCatch, 2025).

La propuesta es un protocolo de detección colaborativa basado en anonymous matching: en lugar de compartir datos personales de usuarios, los actores comparten hashes de identificadores relevantes (número de teléfono, IBAN, correo electrónico) que el sistema de correlación de INCIBE compara con su base de indicadores. Si existe correlación, se notifica a los actores participantes sin revelar los datos originales. Este modelo de privacy-preserving matching es el que utiliza el UK Finance Fraud Managed Service para coordinar la detección de fraude entre entidades financieras en el Reino Unido (UK Finance, 2023), permitiendo participar con riesgo jurídico mínimo porque no implica el intercambio de datos personales en texto claro.

R8 — Sistema de alertas ciudadanas en tiempo real vinculado a la inteligencia antifraude

La distancia temporal entre la activación de una campaña de fraude y el momento en que la información sobre esa campaña llega al ciudadano potencialmente afectado es uno de los factores que más determina el número de víctimas. ENISA ha documentado que los programas de concienciación activa pueden reducir la tasa de clic en campañas de phishing entre un 40% y un 70%, dependiendo del tiempo entre la emisión de la alerta y la recepción del mensaje fraudulento (ENISA, 2023b). El programa ScamShield de Singapur, que combina bloqueo proactivo de números fraudulentos con alertas en tiempo real sobre campañas activas, redujo el número de víctimas por campaña en un 43% en su primer año de funcionamiento (Singapore Police Force, 2024).

El sistema propuesto funcionaría en tres niveles: publicación automatizada en redes sociales de INCIBE y actualización del portal OSI cuando se detecte una campaña activa con más de 10.000 destinatarios (nivel uno); SMS de servicio público a través de los

operadores telco —bajo el mismo marco legal del sistema ES-Alert— cuando la campaña supere los 100.000 destinatarios o incluya vectores de riesgo de daño grave (nivel dos); y comunicados de prensa coordinados con medios de mayor audiencia para campañas de escala excepcional (nivel tres). La implementación del nivel uno es inmediata con los recursos existentes de INCIBE; los niveles dos y tres requieren acuerdos formales con operadores telco y medios, negociables en el marco de la Mesa Técnica Antifraude.

R9 — Incentivos regulatorios para el diseño de servicios con enfoque fraud-resistant UX

Los equipos de producto de las plataformas digitales toman decisiones de diseño que tienen consecuencias directas sobre la vulnerabilidad de sus usuarios al fraude, sin que esas consecuencias sean actualmente visibles ni medidas en sus indicadores de rendimiento. La ratio de conversión en el flujo de registro, el número de pasos en el proceso de cambio de contraseña, o la presencia o ausencia de un aviso al acceder desde un nuevo dispositivo son decisiones de UX que determinan directamente la facilidad con que un atacante puede ejecutar un account takeover. Ninguna de estas decisiones está regulada actualmente en el contexto de las plataformas de contenido en España.

La propuesta es incorporar criterios de fraud-resistant UX en dos instrumentos regulatorios existentes: la evaluación de impacto sobre la protección de datos (DPIA), exigida por el artículo 35 del RGPD (Unión Europea, 2016), incluyendo una sección específica sobre riesgos de diseño de interfaz que facilitan el fraude; y se propone la creación de un sello antifraude, con criterios explícitos de UX como la obligatoriedad de doble factor en cambios de método de pago y un mecanismo visible de reporte de actividad sospechosa. Esta integración es coherente con el enfoque de security by design promovido por el RGPD y con los principios de resiliencia aplicada a plataformas digitales de la Directiva NIS2 (Unión Europea, 2022).

7.4. Educación y resiliencia del usuario

R10 — Modelo de colaboración público-privada estructurada para la concienciación antifraude

La concienciación ciudadana sobre el fraude digital es un bien público que el mercado produce de forma subóptima. Cada plataforma, operador y entidad financiera tiene incentivos para educar a sus propios usuarios sobre el fraude que les afecta directamente, pero nadie tiene incentivos para financiar la educación general sobre el ecosistema del fraude en su conjunto. El resultado es que el ciudadano recibe mensajes fragmentados de diferentes actores, sin una visión afín de cómo se articulan los diferentes tipos de fraude y qué estrategias de autoprotección son generalizables (ENISA, 2023b; Cialdini, 2009).

La solución es un modelo de colaboración estructurada con precedente directo en el programa Cyber Aware del NCSC del Reino Unido. En su edición de 2022-2023, el programa movilizó a más de cuarenta empresas privadas —incluyendo telcos, bancos y plataformas digitales— para difundir mensajes de concienciación comunes a sus respectivas bases de usuarios. El alcance combinado fue de más de 30 millones de personas en el Reino Unido, con un coste por persona alcanzada muy inferior al que habría tenido una campaña equivalente financiada exclusivamente con fondos públicos (NCSC, 2023b). En el contexto español, la Mesa Técnica Antifraude propuesta en R2

sería el foro natural para articular este modelo, acordando un calendario anual de campañas de concienciación y mensajes clave por segmento de audiencia, actualizados trimestralmente en función de las campañas de fraude activas identificadas por la PNIA.

7.5. Marco de métricas e indicadores de seguimiento del sistema antifraude

Las diez recomendaciones estratégicas formuladas en las secciones 7.1 a 7.4 tienen valor operativo únicamente si van acompañadas de un sistema de medición que permita evaluar su grado de implementación, identificar los ámbitos donde el progreso es insuficiente y comparar la evolución del ecosistema antifraude español con referencias internacionales. Esta sección propone un marco de indicadores organizado en torno a los cuatro ejes del modelo PDIP (Prevención, Detección, Intervención, Persecución) y un bloque adicional de indicadores sistémicos y transversales que miden el impacto agregado del fraude sobre la confianza digital y la economía nacional.

El marco está diseñado conforme a tres principios: operacionalidad (cada indicador es calculable con fuentes de datos existentes o accesibles en el corto plazo), comparabilidad internacional (los indicadores se alinean con las métricas promovidas por el GSMA Fraud Management Framework, la metodología de evaluación de ENISA y los reportes de Europol, lo que permite el benchmarking con otros países europeos), y gobernanza clara (cada indicador tiene un actor responsable de su medición y reporte, evitando duplicidades y asimetrías de información entre supervisores).

El marco distingue entre indicadores de resultado —que miden el impacto final del fraude y la eficacia del sistema de defensa— e indicadores de proceso —que miden la capacidad operativa de los actores en cada eje del modelo PDIP—. Esta distinción es relevante para la gestión de la política pública: los indicadores de proceso permiten detectar problemas de implementación antes de que se materialicen en indicadores de resultado negativos, habilitando correcciones proactivas. La relación entre ambos tipos de indicadores refleja, en términos empíricos, la dinámica de coevolución atacante-defensa descrita en el modelo teórico de la Sección 9.2.3.4.

Tabla de indicadores por eje PDIP

Indicador	Definición y forma de cálculo	Fuente de datos principal	Frecuencia / Actor responsable	Eje PDIP
BLOQUE 1 — Prevención				
I-P1. Índice de exposición ciudadana al fraude	% de usuarios que declara haber recibido al menos un intento de fraude (phishing, smishing, deepfake) en los últimos 12 meses. Medido por encuesta representativa a nivel nacional.	Encuesta OSI/INCIBE, Eurobarómetro, Banco de España	Anual / INCIBE	Prevención

I-P2. Tasa de fraude por tipología de plataforma	N.º de incidentes de fraude notificados por cada 100.000 usuarios activos, desagregado por tipo de plataforma (streaming, redes sociales, marketplace, banca digital). Permite comparar exposición relativa entre ecosistemas.	Plataformas (reporting voluntario u obligatorio), FCSE, CNMC	Trimestral / CNMC con INCIBE	Prevención
I-P3. Nivel de concienciación del usuario	% de usuarios capaces de identificar correctamente al menos 3 señales de alerta de fraude (phishing, smishing, perfil falso) en test de conocimiento estandarizado. Complementado con % que activa MFA en sus plataformas.	Encuesta OSI/INCIBE; datos de adopción de MFA de plataformas	Anual / INCIBE – OSI	Prevención
I-P4. Índice de madurez de seguridad por diseño (SbD)	Puntuación compuesta (0-100) que evalúa el grado de adopción de controles de security by design en plataformas de contenido: MFA obligatoria en cambios críticos, autenticación adaptativa, auditoría periódica de APIs, política de divulgación responsable.	Autoevaluación de plataformas, auditorías externas, CNMC / AEPD	Anual / Plataformas + supervisores sectoriales	Prevención
BLOQUE 2 — Detección				
I-D1. Tiempo medio de detección (MTTD)	Tiempo transcurrido (en horas) entre el inicio de una campaña de fraude y la primera alerta técnica generada por el sistema de detección. Métrica crítica: cada hora de retraso amplifica el número de víctimas potenciales.	Sistemas SIEM/UEBA de plataformas y operadores, INCIBE-CERT, registros de incidentes	Continuo; informe mensual / INCIBE + plataformas	Detección

I-D2. Tasa de detección automatizada vs. detección manual	% de incidentes detectados por sistemas automatizados (ML, reglas, UEBA) frente al % detectado por reporte de usuario u otras vías manuales. Un descenso del ratio manual es indicativo de mejora en madurez del sistema de detección.	Registros internos de plataformas, operadores telco y entidades financieras	Mensual / Plataformas y operadores	Detección
I-D3. Tasa de falsos positivos en detección de fraude	% de alertas de fraude generadas que corresponden a actividad legítima de usuarios. Un ratio elevado indica sobrecoste operativo y riesgo de deterioro de la experiencia de usuario; un ratio bajo puede indicar subdetección.	Sistemas de gestión de casos antifraude de plataformas	Mensual / Plataformas	Detección
I-D4. Cobertura de inteligencia compartida intersectorial	% de incidentes de fraude para los que existe correlación de señales de al menos dos sectores distintos (telco + financiero, telco + plataforma, etc.) en la Plataforma Nacional de Inteligencia Antifraude. Mide el grado de integración del ecosistema de detección.	PNIA (Plataforma Nacional de Inteligencia Antifraude / INCIBE)	Mensual / INCIBE	Detección
BLOQUE 3 — Intervención				
I-I1. Tiempo medio de respuesta (MTTR)	Tiempo transcurrido (en horas) entre la primera alerta de una campaña de fraude y la activación de la primera medida de bloqueo o mitigación operativa (bloqueo de dominio, suspensión de número, congelación de cuenta). Indicador clave de la eficacia del protocolo de intervención.	Registros de operadores telco, plataformas, entidades financieras y FCSE	Continuo; informe mensual / INCIBE + coordinadores sectoriales	Intervención

I-I2. Volumen de cuentas comprometidas bloqueadas proactivamente	N.º de cuentas de usuario identificadas como comprometidas (account takeover activo o inminente) que son bloqueadas o puestas en modo de verificación antes de que se complete la fase de monetización del fraude.	Plataformas digitales, entidades financieras	Mensual / Plataformas + Banco de España	Intervención
I-I3. Impacto económico evitado por intervención temprana	Estimación del fraude económico potencial neutralizado mediante intervenciones previas a la monetización. Calculado como (n.º de ataques bloqueados × importe medio histórico por ataque completado). Métrica de alto valor para la justificación de la inversión en el modelo PDIP.	Plataformas, entidades financieras, FCSE (estimaciones forenses)	Trimestral / INCIBE con Banco de España	Intervención
I-I4. Tasa de escalado entre sectores en incidentes coordinados	% de incidentes de fraude que, una vez detectados en un sector, dan lugar a una notificación formal y coordinada a otros sectores afectados dentro de los plazos del protocolo PDIP. Mide la eficacia de los mecanismos de coordinación intersectorial.	Registros del protocolo de coordinación INCIBE / Mesa Técnica Antifraude	Mensual / INCIBE	Intervención
BLOQUE 4 — Persecución				
I-Pr1. % de fraude reportado vs. fraude estimado	Ratio entre el n.º de denuncias formales por fraude digital presentadas ante las FCSE y la estimación del fraude total (calculada a partir de encuestas de victimización). Un ratio bajo indica infradenuncia estructural, lo que limita la eficacia de la persecución y distorsiona la evaluación del riesgo.	FCSE (BIT, UCO), Ministerio del Interior, encuestas de victimización INE/INCIBE	Anual / Ministerio del Interior + INCIBE	Persecución
I-Pr2. Tasa de desarticulación de redes de cuentas mulas	N.º de redes de cuentas mulas identificadas y desarticuladas por las FCSE en relación con el total de redes activas estimadas. Incluye el número de detenidos, el importe de	Ministerio del Interior, Banco de España (informes AML), Europol	Anual / Ministerio del Interior + Banco de España	Persecución

	activos intervenidos y el número de cuentas bloqueadas.			
I-Pr3. Tiempo medio desde denuncia hasta inicio de investigación formal	Tiempo (en días) transcurrido entre la presentación de una denuncia por fraude digital y la apertura de diligencias de investigación formal por parte de las FCSE o la Fiscalía. Indicador de la capacidad de respuesta del sistema judicial y policial.	FCSE, Ministerio de Justicia, Fiscalía de Criminalidad Informática	Anual / Ministerio de Justicia	Persecución
I-Pr4. Tasa de recuperación de activos en fraude digital	% del importe total defraudado en incidentes investigados que es recuperado o restituido a las víctimas mediante resolución judicial o acuerdo. Incluye activos en criptomonedas. Métrica de difícil medición pero de alto valor para la evaluación de la eficacia del sistema penal.	Ministerio de Justicia, FCSE, plataformas de criptomonedas reguladas	Anual / Ministerio de Justicia + Banco de España	Persecución
BLOQUE 5 — Indicadores sistémicos y transversales				
I-S1. Índice de confianza digital del ciudadano	Puntuación compuesta (0-100) que mide la percepción de seguridad del ciudadano en el uso de servicios digitales (comercio electrónico, banca digital, plataformas de contenido). Se desagrega por grupo de edad, nivel de formación y tipo de servicio. Permite evaluar el impacto del fraude sobre la adopción de la economía digital.	Encuesta representativa (INE/ONTSI/INCIBE), Eurobarómetro Digital	Anual / INCIBE + ONTSI (Observatorio Nacional de Tecnología y Sociedad)	Sistémico
I-S2. Índice de madurez del ecosistema antifraude nacional	Evaluación compuesta (0-5) del grado de desarrollo del ecosistema antifraude español en cinco dimensiones: inteligencia compartida, coordinación sectorial, capacidades técnicas de detección, marco regulatorio y concienciación ciudadana. Permite comparación	Autoevaluación coordinada por INCIBE con metodología GSMA Fraud Management Maturity Model	Bienal / INCIBE con CNMC, Banco de España y AEPD	Sistémico

	internacional con metodología GSMA/ENISA.			
I-S3. Coste total del fraude digital como % del PIB digital	Ratio entre la estimación del impacto económico total del fraude digital (pérdidas directas + costes de gestión + impacto sobre la confianza) y el PIB digital español (economía digital como % del PIB total). Permite evaluar la dimensión macroeconómica del fenómeno y la eficacia de las medidas de mitigación en el tiempo.	Banco de España, Visa España, CNMC, informes anuales ENISA/Europol	Anual / Banco de España + CNMC	Sistémico

Tabla 1 Marco de indicadores de seguimiento del sistema antifraude nacional, organizados por eje del modelo PDIP. Elaboración propia a partir de GSMA (2025), CFCA (2023), Europol (2023) e INCIBE.

7.5.1 Consideraciones de gobernanza del sistema de medición

La implementación del marco de indicadores requiere resolver tres cuestiones de gobernanza que condicionan su viabilidad práctica.

Propiedad y reporte. Se propone que INCIBE asuma la responsabilidad de la agregación, validación y publicación anual de los indicadores sistémicos (bloque I-S) y de los indicadores de los ejes de detección e intervención que requieren correlación intersectorial (I-D4, I-I1, I-I4). Los indicadores de los ejes de prevención y persecución recaerían bajo la responsabilidad primaria de las plataformas y operadores (con supervisión de CNMC y Banco de España) y del Ministerio del Interior/Ministerio de Justicia, respectivamente, con reporte anual consolidado a INCIBE para su integración en el cuadro de mando nacional.

Confidencialidad y agregación. Varios indicadores (especialmente los de detección y tasa de fraude por plataforma) implican el tratamiento de datos operacionales sensibles cuya divulgación podría generar riesgo reputacional o proporcionar inteligencia a los actores fraudulentos. El marco de gobernanza debe establecer niveles de clasificación de la información análogos a los utilizados en el intercambio de inteligencia de amenazas: un nivel de acceso restringido para los actores del ecosistema participantes en la Mesa Técnica Antifraude, y un nivel de publicación pública que agregue y anonimice los datos antes de su difusión externa.

Línea de base y metas. La implementación del marco de indicadores debe comenzar con una fase de establecimiento de la línea de base (baseline) durante el primer año, en la que se calibran los métodos de medición, se acuerdan las definiciones operativas con los actores responsables y se recopilan los primeros datos comparables. Las metas cuantitativas para cada indicador —que no se establecen en este trabajo dada la ausencia de una línea de base oficial— deben fijarse en el contexto de la Mesa Técnica Antifraude

propuesta en la Recomendación R2, teniendo como referencia los valores medios europeos disponibles en los reportes anuales de Europol, ENISA y el GSMA.

En conjunto, el marco de indicadores propuesto cierra el ciclo de gestión del sistema antifraude nacional: las recomendaciones del Capítulo 7 definen qué hay que hacer; el modelo PDIP de la Sección 9.1 define cómo hay que organizarlo; y este sistema de métricas define cómo saber si se está haciendo bien y en qué dirección hay que corregir el rumbo. Sin este tercer elemento, el riesgo de las políticas antifraude —como el de cualquier política pública— es el de convertirse en declaraciones de intención sin capacidad de rendición de cuentas ni aprendizaje institucional.

7.6. Rol de INCIBE en el ecosistema antifraude nacional

El Instituto Nacional de Ciberseguridad (INCIBE) constituye el organismo de referencia en materia de ciberseguridad para ciudadanos, empresas y operadores de servicios esenciales en España, con el mandato institucional de fortalecer la resiliencia digital del país mediante la prevención, detección y respuesta ante ciberincidentes. En el contexto del fraude digital en plataformas de contenido, el análisis desarrollado en los capítulos anteriores revela un escenario en el que la respuesta eficaz no puede depender exclusivamente de las capacidades individuales de cada actor sectorial, sino que requiere un nodo de coordinación con autoridad técnica, visibilidad transversal y capacidad de articulación público-privada. INCIBE reúne estas condiciones de forma única en el contexto español.

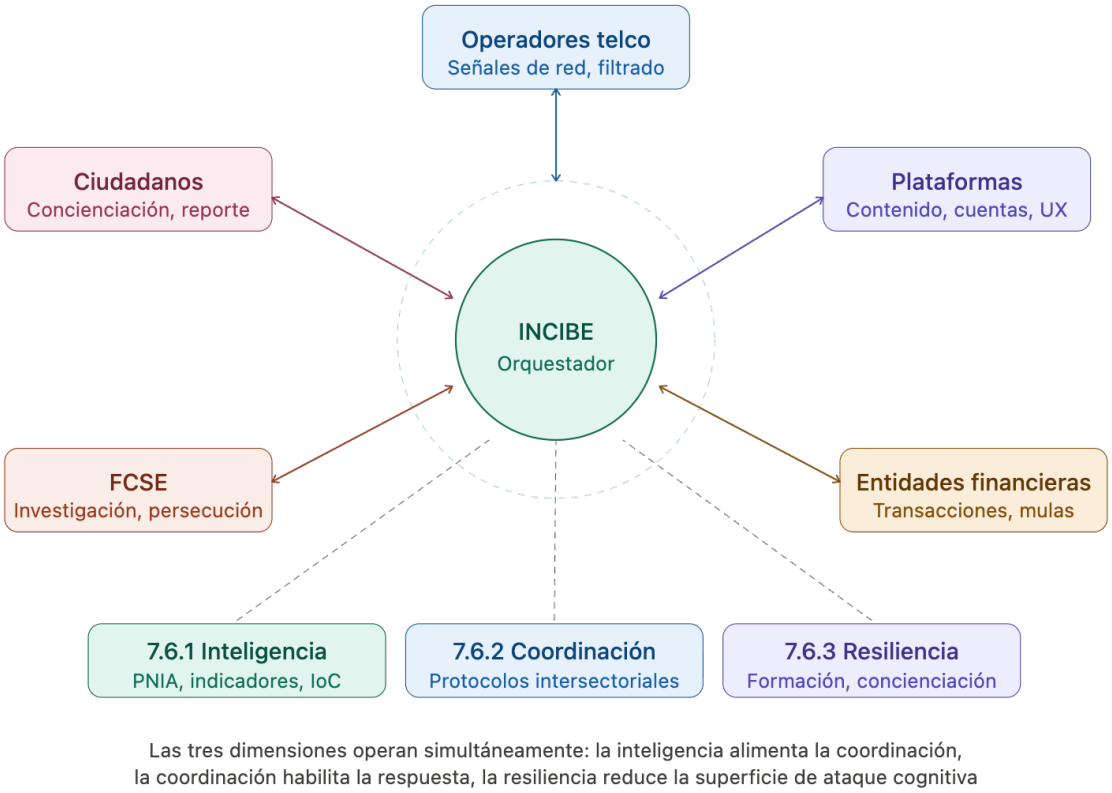

Figura 6 INCIBE como orquestador del sistema antifraude nacional: modelo hub-and-spoke con tres dimensiones funcionales (inteligencia, coordinación, resiliencia) y cinco categorías de actores del ecosistema. Elaboración propia.

La naturaleza del fraude digital descrita en este trabajo —industrializada, adaptativa, convergente entre sectores telco, financiero y de plataformas— demanda una evolución del rol de INCIBE desde el modelo clásico de respuesta a incidentes hacia una función de orquestación estratégica del sistema antifraude nacional. Esta evolución no implica sustituir las capacidades sectoriales existentes, sino articularlas en torno a una arquitectura de inteligencia, coordinación y resiliencia que multiplique su eficacia conjunta.

7.6.1 Nodo de inteligencia nacional sobre fraude digital

La primera dimensión del rol propuesto sitúa a INCIBE como nodo central de agregación, análisis y diseminación de inteligencia sobre fraude digital en plataformas de contenido. En la actualidad, los operadores de telecomunicaciones, las entidades financieras y las plataformas digitales generan de forma independiente señales de fraude —indicadores de compromiso, patrones de abuso, campañas de phishing activas, identificadores de infraestructura criminal— que raramente se ponen en común de manera estructurada y en tiempo real. Esta fragmentación reduce la capacidad de detección temprana y permite que ataques ya conocidos por un actor afecten a otros que carecen de la señal.

En este marco, INCIBE puede asumir el papel de operador del repositorio nacional de inteligencia antifraude, articulado en torno a tres funciones: la ingesta automatizada y normalizada de indicadores de fraude procedentes de múltiples sectores, el análisis de patrones emergentes mediante técnicas de correlación y aprendizaje automático, y la generación y diseminación de alertas tempranas hacia los actores del ecosistema antes de que las campañas fraudulentas alcancen escala. Este modelo es coherente con la función que el Centro Europeo de Cibercrimen (EC3) de Europol ejerce a escala supranacional, y con la experiencia de iniciativas sectoriales como el Fraud Intelligence Sharing Service (FISS) impulsado por el GSMA, cuyas metodologías y estándares de intercambio pueden adoptarse como referencia técnica para la arquitectura nacional.

La viabilidad de este nodo requiere dos condiciones estructurales. La primera es la existencia de un marco jurídico que habilite el intercambio de datos de fraude entre sectores —incluidos datos de carácter personal tratados con base en el interés legítimo de prevención— con garantías de confidencialidad y proporcionalidad compatibles con el RGPD. La segunda es la adopción de estándares técnicos comunes de taxonomía y formato para la comunicación de indicadores, como los que promueven MISP (Malware Information Sharing Platform) o STIX/TAXII, que permitan la interoperabilidad entre sistemas heterogéneos. INCIBE dispone del capital institucional y la posición regulatoria para impulsar ambas condiciones en el marco de la Estrategia Nacional de Ciberseguridad.

7.6.2 Plataforma de coordinación sectorial intersectorial

La segunda dimensión del rol de INCIBE en el ecosistema antifraude es la coordinación operativa entre actores heterogéneos que, aunque afectados por el mismo fenómeno, operan bajo marcos regulatorios, incentivos y capacidades técnicas distintos. El análisis de la infraestructura criminal del fraude (Sección 4.7) evidencia que los ataques exitosos explotan precisamente las discontinuidades entre el perímetro de responsabilidad de cada actor: el operador telco detecta la campaña de smishing pero no tiene visibilidad sobre la transacción financiera fraudulenta que la remata; la entidad financiera identifica el cargo no autorizado pero desconoce el vector de captación que lo originó; la plataforma digital

observa el abuso de credenciales pero no dispone de la señal de riesgo de red que lo precede.

La superación de estas discontinuidades requiere un modelo de coordinación sectorial con tres elementos: una mesa técnica permanente que integre a representantes de operadores de telecomunicaciones (Telefónica, Vodafone, Orange, MásMóvil), entidades financieras reguladas por el Banco de España, plataformas de contenido digital con presencia relevante en el mercado español, y agencias de aplicación de la ley (Policía Nacional, Guardia Civil, Fiscalía de Criminalidad Informática); un protocolo de respuesta coordinada ante incidentes de fraude masivo, con plazos, roles y canales de comunicación predefinidos; y un catálogo de estándares de autenticación e intercambio de datos que establezca los requisitos mínimos de interoperabilidad para la prevención del fraude en servicios digitales de alto riesgo.

Esta función de coordinación no es nueva para INCIBE, que ya opera como punto de contacto nacional en el marco de la Directiva NIS2 y participa en redes europeas de CSIRTs. La ampliación hacia el dominio del fraude en plataformas de contenido representa una extensión natural de este mandato, coherente con la convergencia entre ciberseguridad y fraude financiero que caracteriza el panorama de amenazas en 2025-2026.

7.6.3 Motor de concienciación y resiliencia ciudadana

La tercera dimensión del rol de INCIBE en el ecosistema antifraude aborda el vector más difícil de mitigar desde la perspectiva técnica: la vulnerabilidad cognitiva del usuario. El análisis de los mecanismos de ingeniería social (Sección 4.1) y el impacto sobre los usuarios (Sección 1.2.2) demuestran que la eficacia del fraude digital descansa en gran medida en la explotación de sesgos conductuales —urgencia, autoridad, escasez, reciprocidad— que no pueden neutralizarse únicamente mediante controles tecnológicos. La educación digital del usuario es, en este sentido, un complemento insustituible de los sistemas de detección y autenticación.

INCIBE ya opera el portal OSI (Oficina de Seguridad del Internauta) y desarrolla programas de concienciación dirigidos a distintos segmentos de la población, incluyendo menores, mayores y empresas. La propuesta en este contexto es la evolución de estos programas hacia un modelo de resiliencia activa frente al fraude en plataformas de contenido, estructurado en tres niveles: campañas de alerta inmediata sobre campañas de fraude activas, vinculadas al sistema de inteligencia descrito en el epígrafe anterior y con capacidad de difusión a través de canales de alta penetración (SMS institucional, redes sociales, colaboración con operadores telco); programas de formación continua orientados a grupos de alta vulnerabilidad, con especial atención a la detección de deepfakes, el reconocimiento de patrones de phishing y la gestión segura de credenciales en plataformas de contenido; y una iniciativa de certificación o sello de confianza antifraude para plataformas digitales que adopten estándares de diseño seguro, autenticación reforzada y transparencia en la gestión de fraude, comparable al modelo de certificación ENS en el ámbito de la administración pública.

En conjunto, las tres dimensiones descritas —inteligencia, coordinación y resiliencia— configuran un modelo de INCIBE como orquestador del sistema antifraude nacional, coherente con la naturaleza sistémica del fenómeno analizado en este trabajo y con la posición institucional que INCIBE ocupa en el ecosistema de ciberseguridad español. La

implementación gradual de este modelo, comenzando por la mesa técnica de coordinación sectorial y el protocolo de intercambio de inteligencia, representaría un salto cualitativo en la capacidad colectiva de respuesta al fraude digital en España.

8. Casos de fraude end-to-end: anatomía de la cadena de ataque

Este capítulo presenta tres casos end-to-end construidos a partir de la síntesis de incidentes documentados por el GSMA Fraud and Security Group en sus sesiones FASG#33 (2025) y FASG#34 (2026), los informes anuales de Europol (2023b), los datos del Ministerio del Interior español (2026), y los análisis de Kaspersky (2025, 2026) y McAfee (2024, 2026). Cada caso está tipificado —no atribuido a actores específicos identificables— pero refleja patrones operativos recurrentes en el ecosistema de fraude digital europeo y español, documentados en las fuentes citadas. Las estadísticas específicas de cada caso se atribuyen a los rangos publicados por estas fuentes primarias, con indicación explícita cuando se trata de estimaciones derivadas de esos rangos. El propósito es ilustrar los puntos de la cadena en los que una intervención oportuna habría podido interrumpirla, conectando la anatomía del fraude con las recomendaciones operativas del capítulo anterior.

8.1. Caso A — El ciclo completo del fraude de suscripción: de la campaña de smishing a la red de mulas

En enero de 2026, los sistemas de monitorización de un operador de telecomunicaciones español detectaron un incremento anómalo del tráfico SMS saliente desde un grupo de números de origen extranjero que habían activado SIM españolas en los últimos quince días. Los mensajes, enviados a una tasa de aproximadamente 8.000 por hora —consistente con la capacidad operativa de los SMS blasters documentados por el GSMA en el contexto europeo (GSMA FASG#34, 2026)—, contenían textos que simulaban notificaciones de una plataforma de streaming de vídeo ampliamente utilizada en España, informando al receptor de un problema con el método de pago de su suscripción e instándole a resolver la incidencia a través de un enlace acortado. El nombre del remitente alfanumérico en el encabezado del SMS coincidía con el identificador oficial de la plataforma, porque los atacantes habían utilizado un SMS blaster con capacidad de spoofing del Sender ID —una vulnerabilidad que persiste en la red española al no estar implementado el Sender ID Registry en todos los operadores (GSMA FASG#34, 2026; ComReg, 2025).

El enlace redirigía a un dominio registrado tres días antes con el patrón 'nombre-de-la-plataforma-verificacion.com', alojado en un servidor de hosting dedicado en un país del Este de Europa. La página reproducía fielmente la interfaz de inicio de sesión de la plataforma legítima y solicitaba credenciales de acceso y, en un segundo paso, los datos completos de la tarjeta de crédito. La sofisticación técnica incluía un proxy transparente hacia la plataforma real, de forma que el usuario veía confirmado que sus credenciales eran correctas y que el problema de pago había sido resuelto, sin percibir ninguna señal de alerta. Esta técnica de reverse proxy phishing está documentada en el informe de amenazas avanzadas de Europol (2023b) como una de las variantes de mayor eficacia en las campañas de credential harvesting contra plataformas de streaming.

Durante las primeras cuatro horas, la campaña alcanzó a un volumen de usuarios consistente con los rangos documentados para campañas de smishing de escala media en el ecosistema europeo —entre 50.000 y 500.000 destinatarios, según el GSMA FASG#34

(2026). La tasa de clic, estimada en el rango del 3% al 8% documentado por el GSMA para campañas de smishing con suplantación de plataformas de contenido (GSMA FASG#33, 2025), generó varios miles de visitas a la página fraudulenta, de las cuales aproximadamente el 20% completaron el formulario con credenciales y datos de tarjeta —porcentaje consistente con el análisis de conversión de páginas clon publicado por Kaspersky (2026). El sistema de clasificación automatizado del panel de control de los atacantes, accesible desde Telegram, etiquetó como 'premium' los registros cuyas credenciales fueron verificadas en tiempo real contra la plataforma y cuya tarjeta pasó una microtransacción de prueba. En cuestión de horas, estos registros comenzaron a circular en un canal privado de Telegram a precios de entre 8 y 15 euros por registro —rango documentado en el análisis de mercados negros de credenciales de Europol (2023b).

Paralelamente, los atacantes utilizaron las credenciales capturadas para ejecutar ataques de credential stuffing contra otras plataformas. La tasa de éxito del stuffing en el contexto europeo oscila entre el 0,5% y el 2%, según el análisis de Europol (2023b), lo que en campañas de escala genera centenares de cuentas adicionales comprometidas por cada campaña inicial. Las transferencias hacia cuentas mulas se realizaron en importes inferiores a 950 euros —justo por debajo del umbral habitual de verificación adicional documentado por el Banco de España (2025)—, y los fondos fueron convertidos a criptomonedas en un plazo inferior a cuatro horas por la red de mulas, en un patrón operativo consistente con el descrito por el Ministerio del Interior (2026) en su análisis de redes de blanqueo desarticuladas en España.

La cadena completa —desde el envío del primer SMS hasta la conversión a criptomonedas— se completó en menos de dieciséis horas. Las señales del ataque estuvieron disponibles en distintos sistemas durante toda la cadena, pero ningún actor disponía de la visión completa ni del protocolo para correlacionarlas y activar una respuesta conjunta. El operador telco detectó el patrón de envío masivo pero no tenía manera automatizada de notificarlo a la plataforma de streaming ni al sistema financiero. La plataforma detectó accesos anómalos desde nuevas IPs, pero cuando intentó contactar con el operador, la comunicación fue a través de canales de atención al cliente no diseñados para intercambio de inteligencia en tiempo real. Los bancos detectaron las microtransacciones de prueba, pero las gestionaron como fraude de tarjeta ordinario, sin relacionarlo con el vector de smishing original.

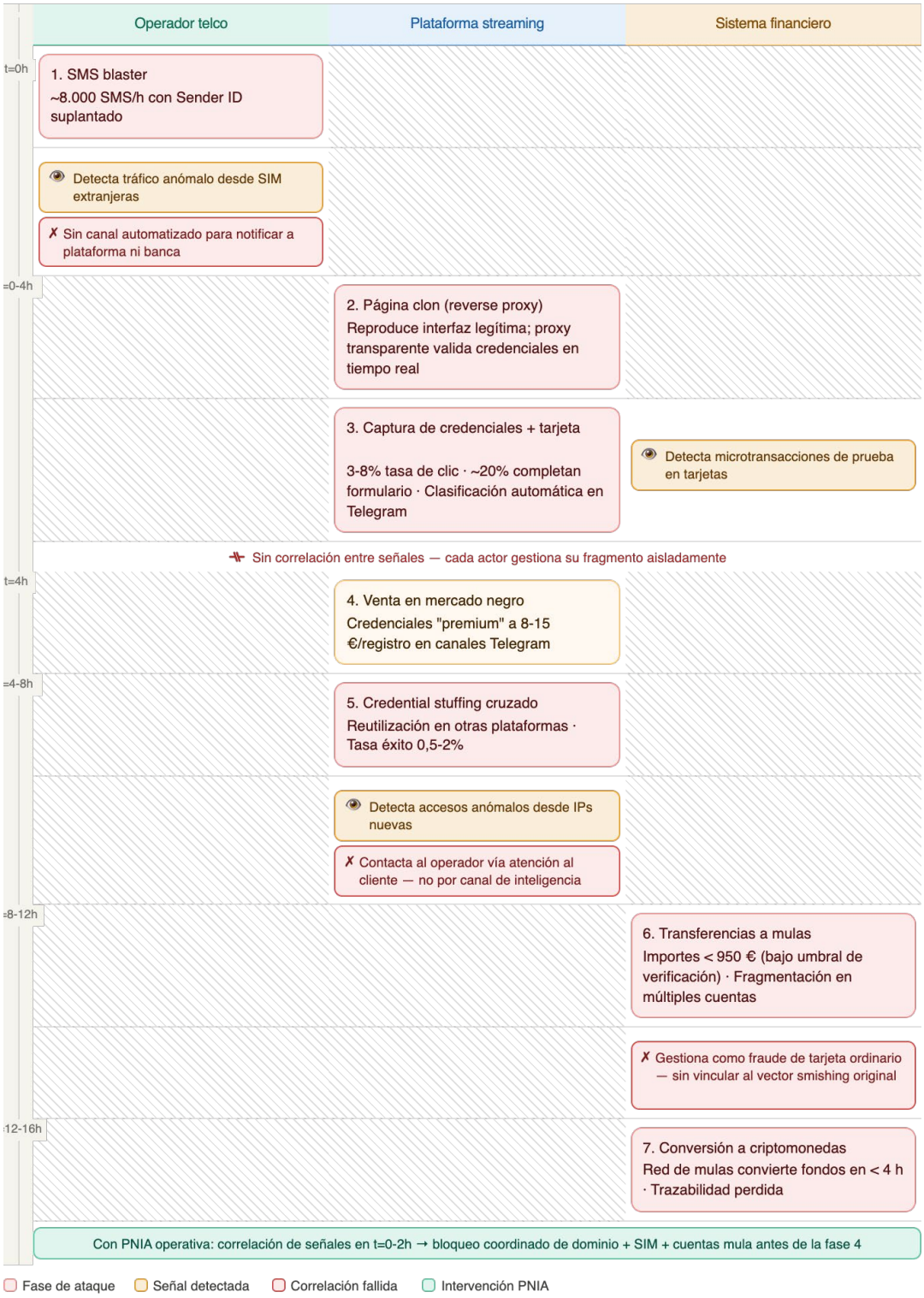

Figura 7 Caso A — Cadena de ataque del fraude de suscripción vía smishing: diagrama swimlane con tres actores (operador telco, plataforma de streaming, sistema financiero), señales detectadas, fallos de correlación y punto de intervención PNIA. Elaboración propia.

Este escenario de fragmentación operativa es precisamente el que el GSMA Fraud and Security Group identifica como el principal factor amplificador del daño en ataques de cadena múltiple (GSMA FASG#33, 2025).

8.2. Caso B — La estafa de inversión mediante deepfake: cuando la víctima colabora voluntariamente

El perfil demográfico de las víctimas de las estafas de inversión documentadas en España en 2025 sorprende a quien asume que el fraude digital afecta principalmente a usuarios sin formación digital. Los datos del Ministerio del Interior (2026) y del Banco de España (2025) sobre denuncias por estafas de inversión online revelan que la víctima típica de este vector tiene entre 48 y 62 años, nivel de ingresos medio-alto, experiencia previa con productos financieros y un nivel de uso digital suficientemente alto como para sentirse cómodo realizando transferencias bancarias online. Este perfil coincide con el documentado por la Global Initiative Against Transnational Organized Crime (2026) en su análisis global de scam centres, que describe cómo los operadores de fraude de inversión diseñan sus campañas explícitamente para captar usuarios con capacidad de ahorro, no usuarios sin formación, porque su mayor poder adquisitivo maximiza el rendimiento por víctima.

El ciclo típico documentado en los informes del Banco de España y el Ministerio del Interior (2025-2026) comienza con un anuncio en las plataformas de publicidad de redes sociales que muestra un vídeo en el que una figura pública reconocible en España describe una oportunidad de inversión con rentabilidades garantizadas. El vídeo es un deepfake generado con herramientas de síntesis facial y de voz disponibles comercialmente. El ISMS Forum (2026) documenta que en 2025 el volumen de deepfakes generados con fines de fraude financiero se estimó en torno a los 8 millones de archivos, duplicándose aproximadamente cada seis meses desde 2024, y que las pérdidas asociadas a fraudes financieros mediante deepfakes superaron los 1.500 millones de dólares a escala global. Las señales de manipulación del deepfake —artefactos en los bordes del rostro, ligera desincronización labial, aplanamiento de la textura de la piel— son detectables con herramientas especializadas pero pasan desapercibidas para el ojo no entrenado (ISMS Forum, 2026; Keepnet Labs, 2026).

El anuncio redirige a una página de inversión fraudulenta desde la que un agente del centro de llamadas —localizado frecuentemente en el sudeste asiático, según el informe de la Global Initiative Against Transnational Organized Crime (2026)— contacta al usuario en pocas horas. El proceso de 'engorde' de la víctima —denominado pig butchering en la literatura especializada (Global Initiative Against Transnational Organized Crime, 2026)— puede durar semanas o meses, con el usuario invirtiendo cantidades progresivamente mayores a medida que su confianza se consolida ante rentabilidades ficticias mostradas en el panel de control. Las pérdidas medias documentadas por el Banco de España para las víctimas que superan la primera transferencia se sitúan en importes que en muchos casos exceden los 10.000 euros, con casos extremos en los que la víctima liquidó activos reales para financiar la inversión (Banco de España, 2025). El momento del 'golpe' llega cuando el usuario intenta retirar sus fondos y la plataforma le exige el pago de impuestos o comisiones ficticias, tras lo cual desaparece sin dejar rastro accesible.

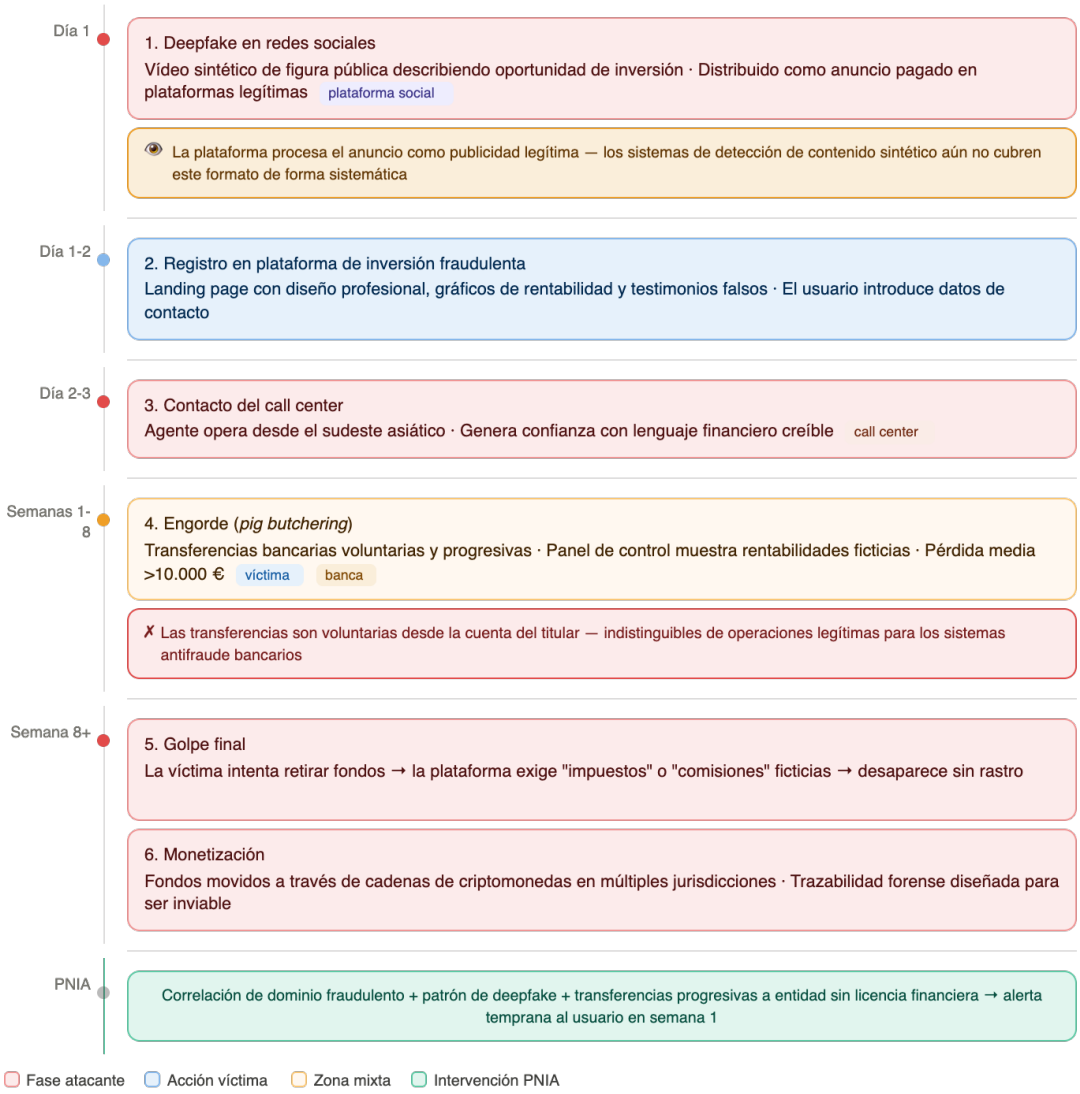

Figura 8 Caso B — Cadena de ataque de la estafa de inversión mediante deepfake (pig butchering): línea temporal desde la distribución del anuncio sintético hasta la monetización en criptomonedas. Elaboración propia a partir de Banco de España (2025), Global Initiative Against Transnational Organized Crime (2026) e ISMS Forum (2026).

La dificultad de detección y respuesta ante este vector es mayor que en el Caso A por varias razones. La víctima realiza las transferencias de forma voluntaria desde su propia cuenta bancaria, lo que las hace indistinguibles de una transferencia legítima para los sistemas antifraude bancarios. El deepfake se distribuye a través de los servidores de publicidad de plataformas legítimas, lo que dificulta su identificación y eliminación. Los sistemas de detección automatizada de contenido sintético se encuentran todavía en fase de desarrollo y despliegue, con brechas de eficacia que el DARPA MediFor programme identificó como estructurales: los modelos de generación evolucionan más rápido que los de detección (DARPA, 2023). El centro de llamadas opera desde jurisdicciones con baja cooperación judicial, y los fondos se mueven a través de cadenas de criptomonedas diseñadas para dificultar la trazabilidad forense (Global Initiative Against Transnational Organized Crime, 2026).

8.3. Caso C — SpyLoan: cuando la app es el arma

El tercer caso ilustra el fraude mediante aplicaciones maliciosas de préstamos rápidos, denominado SpyLoan por la categorización de Kaspersky en su análisis de amenazas móviles de 2023 (Kaspersky, 2023). Este vector es menos visible mediáticamente que el phishing y los deepfakes, pero con un impacto sobre las víctimas —en términos de daño psicológico— que en muchos casos supera el perjuicio económico directo.

La aplicación llega al dispositivo de la víctima a través de tiendas de apps oficiales o alternativas, con la apariencia de un servicio de microcrédito rápido, decenas de valoraciones positivas (generadas mediante cuentas falsas) y diseño visual equivalente al de las apps de fintech legítimas. El proceso de solicitud de préstamo incluye una solicitud de permisos extensivos —contactos, galería, micrófono, localización— justificados con explicaciones plausibles. Kaspersky documentó en su análisis de amenazas móviles del segundo trimestre de 2025 que las variantes de SpyLoan activas en Europa extraen sistemáticamente la agenda de contactos completa, el historial de llamadas, las fotografías almacenadas y la geolocalización en tiempo real, transmitiendo estos datos a servidores externos mediante conexiones HTTPS cifradas (Kaspersky, 2025). McAfee, en su análisis publicado en noviembre de 2024, identificó más de sesenta variantes activas de SpyLoan dirigidas a usuarios europeos, con un tiempo medio de disponibilidad en tiendas oficiales antes de su retirada de 47 días (McAfee, 2024).

El préstamo se aprueba efectivamente, pero con condiciones abusivas —tipos de interés anuales superiores al 500% y plazos de devolución de siete días— que la víctima frecuentemente no lee con detenimiento. La fase de extorsión comienza el día ocho, cuando la app envía mensajes a los contactos de la agenda de la víctima afirmando que es una deudora morosa, en algunos casos acompañados de fotografías tomadas de su galería. Este mecanismo de extorsión mediante daño reputacional, documentado en detalle por McAfee (2024) y analizado en el contexto español por el Centro de Delitos Informáticos del Ministerio del Interior (2026), es brutalmente eficaz: muchas víctimas pagan importes muy superiores al préstamo original para detener el envío de mensajes a sus contactos, sin reportar el incidente precisamente por temor a que el reporte amplíe el daño reputacional.

El análisis de casos reportados al Centro de Denuncias de Delitos Informáticos del Ministerio del Interior (2026) revela más de 3.400 denuncias relacionadas con este vector en España durante 2024-2025, con un importe medio de pérdida que las fuentes del sector sitúan en torno a los 1.800 euros por víctima, aunque la cifra real se estima entre dos y cinco veces mayor debido a la infradenuncia motivada por vergüenza y miedo reputacional. Las apps implicadas habían sido descargadas, en la mayoría de los casos, entre 50.000 y 200.000 veces antes de ser retiradas de las tiendas. La respuesta eficaz a este vector requiere actuaciones simultáneas en tres frentes: refuerzo de los procesos de revisión de las tiendas de apps para la categoría de aplicaciones financieras (una restricción que Google comenzó a implementar en Play Store en 2024 pero cuya aplicación práctica sigue siendo insuficiente, según McAfee (2024)); integración de la capacidad de los operadores telco para identificar patrones de exfiltración de datos con el nodo de inteligencia de INCIBE; y refuerzo de la cooperación judicial internacional con las jurisdicciones donde se alojan los servidores, con India y Nigeria como principales países de origen documentados (McAfee, 2024; Kaspersky, 2025).

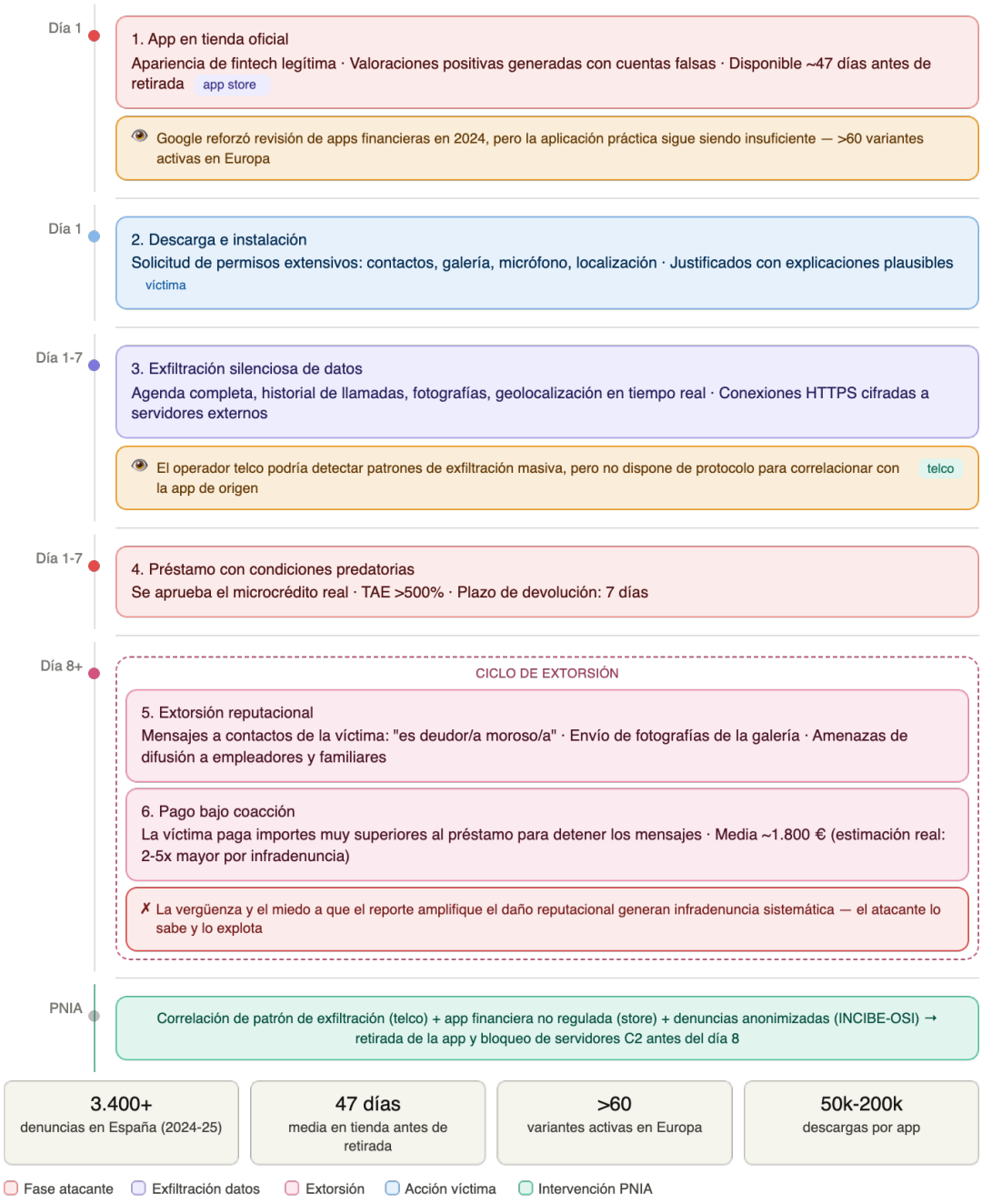

Figura 9 Caso C — Cadena de ataque SpyLoan: desde la descarga de la app hasta el ciclo de extorsión reputacional, con indicadores de escala del fenómeno en España. Elaboración propia a partir de Kaspersky (2025), McAfee (2024) y Ministerio del Interior

8.4. Patrones comunes identificados

Los tres casos analizados comparten cuatro patrones transversales.

- Primero, la velocidad: en los tres casos, el ciclo de daño se completa antes de que los sistemas de respuesta puedan activarse de forma coordinada.
- Segundo, la invisibilidad entre sectores: ninguno de los actores afectados dispone de la señal completa que permitiría identificar el ataque en sus fases iniciales, una

limitación estructural documentada por el GSMA como el principal factor de escalado del daño en ataques de cadena múltiple (GSMA FASG#33, 2025).

- Tercero, la explotación de la confianza en plataformas legítimas: en los tres casos, la credibilidad del ataque descansa sobre la apariencia o el canal de distribución de servicios legítimos reconocidos por la víctima (Cialdini, 2009; Kaspersky, 2026).
- Cuarto, la dificultad de la persecución: en los tres casos, la infraestructura criminal está distribuida en múltiples jurisdicciones y opera con velocidades de despliegue que superan los tiempos de respuesta de los sistemas legales y policiales (Europol, 2023b; Global Initiative Against Transnational Organized Crime, 2026).

Estos cuatro patrones son los que configuran el mapa de prioridades operativas del sistema antifraude nacional propuesto en este trabajo.

9. Modelos conceptuales: marco operativo y modelo socio-técnico del fraude

Este capítulo presenta el modelo de actuación propuesto para la mitigación del fraude digital, articulado en cuatro ejes (prevención, detección, intervención y persecución), seguido del modelo socio-técnico formal del fraude en plataformas de contenido (MSFPC), que formaliza las relaciones entre constructos y desarrolla hipótesis contrastables.

9.1. Modelo de actuación propuesto para la mitigación del fraude digital

Las cuatro implicaciones estratégicas descritas en la sección anterior convergen en una conclusión operativa: la respuesta eficaz al fraude digital en plataformas de contenido requiere un modelo de actuación que integre múltiples capas, múltiples actores y múltiples horizontes temporales. Los enfoques sectoriales, reactivos o puramente tecnológicos han demostrado ser insuficientes frente a un fenómeno que es, en esencia, un sistema sociotécnico adaptativo. El modelo que se propone a continuación se estructura en cuatro ejes secuencialmente articulados — Prevención, Detección, Intervención y Persecución (PDIP) — que cubren el ciclo completo de la respuesta al fraude, desde la reducción de la superficie de ataque hasta la persecución y desarticulación de las organizaciones criminales.

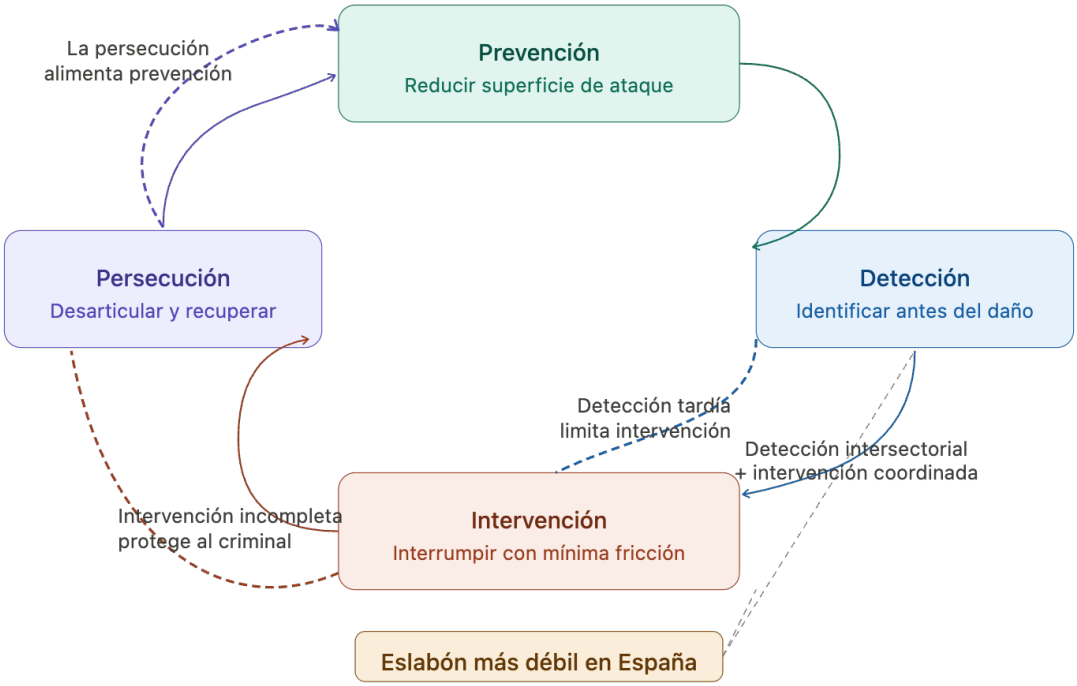

Figura 10 Ciclo PDIP (Prevención, Detección, Intervención, Persecución): ejes del modelo operativo con interdependencias, retroalimentaciones y eslabón más débil en el contexto español. Elaboración propia a partir de GSMA (2025) y CFCA (2023).

Este modelo no es original en su nomenclatura: adopta la estructura de ciclo de vida de la gestión del fraude promovida por el GSMA Fraud Management Framework (GSMA,

2025) y la extiende con las especificidades del contexto español y las dimensiones de coordinación intersectorial que las secciones precedentes han identificado como críticas. Su novedad reside en la integración de las cuatro dimensiones en un marco coherente y priorizado para el contexto nacional, con actores asignados y puntos de intervención concretos derivados del análisis de los vectores de ataque documentados en este trabajo.

Eje 1 — Prevención: reducir la superficie de ataque y fortalecer la resiliencia

La prevención actúa antes de que se produzca el ataque y tiene como objetivo reducir las vulnerabilidades que los actores fraudulentos explotan. En el contexto del fraude en plataformas de contenido, las vulnerabilidades relevantes son de tres tipos: cognitivas (la susceptibilidad del usuario a la ingeniería social), técnicas (debilidades en la autenticación, el diseño de plataformas y la infraestructura de red) y organizativas (ausencia de protocolos de respuesta coordinada entre actores).

La prevención cognitiva requiere programas de educación digital dirigidos específicamente a los segmentos de mayor vulnerabilidad —mayores, personas con menor formación digital, usuarios de plataformas de alto riesgo como aplicaciones de citas o servicios de inversión— con contenidos adaptados a los vectores de fraude prevalentes en el momento. La referencia de los materiales de la Oficina de Seguridad del Internauta (OSI) de INCIBE es válida, pero su alcance y su capacidad de actualización en tiempo real a medida que emergen nuevas campañas fraudulentas necesitan refuerzo. Los operadores telco y las plataformas digitales, con acceso directo a millones de usuarios, son actores clave en la cadena de distribución de estos contenidos preventivos.

La prevención técnica abarca dos dimensiones complementarias. La primera es el diseño seguro de plataformas (security by design), que implica incorporar controles antifraude en la arquitectura de los servicios desde su concepción, no como capa añadida posterior: gestión del ciclo de vida de credenciales, verificación de identidad en el onboarding, autenticación adaptativa basada en riesgo, y auditoría continua de las superficies de ataque expuestas por APIs y sistemas de pago. La segunda es la reducción de la fricción explotable en la experiencia de usuario (fraud-resistant UX): los diseños que maximizan la conversión y minimizan la fricción en el acceso y en los flujos de pago generan vulnerabilidades que los atacantes explotan sistemáticamente, como documentan los casos de credential stuffing y account takeover analizados en la Sección 3.1. El equilibrio entre usabilidad y seguridad es un problema de diseño con implicaciones regulatorias que los supervisores deben empezar a abordar explícitamente.

La prevención organizativa implica que las empresas, plataformas y administraciones públicas dispongan de políticas, procedimientos y capacidades internas de gestión del riesgo de fraude. Esto incluye la formación periódica de los equipos de atención al cliente —primer punto de contacto con víctimas potenciales—, la existencia de protocolos de respuesta ante incidentes de fraude masivo, y la participación activa en los mecanismos de inteligencia compartida que se desarrollen en el marco de las estructuras de coordinación propuestas en la Sección 4.9.1.

Eje 2 — Detección: identificar el fraude antes de que cause daño irreversible

La detección es el eje más técnicamente intensivo del modelo y el que mayor evolución ha experimentado en los últimos años gracias al aprendizaje automático y el análisis de comportamiento. El objetivo es reducir el tiempo entre la primera señal de una campaña

de fraude y la activación de la respuesta, minimizando el número de víctimas y el importe del daño.

Los sistemas de detección de anomalías y análisis de comportamiento de usuarios y entidades (UEBA) descritos en la Sección 6.2 del documento —en el capítulo de mecanismos de detección y prevención— constituyen el núcleo técnico de este eje. Sin embargo, su eficacia está condicionada a la calidad y diversidad de los datos con los que se entrenan: un modelo entrenado exclusivamente con señales de una plataforma de streaming tiene una capacidad de detección limitada frente a ataques que comienzan en la red telco y se materializan en el sistema financiero. La correlación de señales entre sectores —lo que el GSMA denomina cross-industry intelligence sharing— es el factor diferencial que permite detectar campañas de fraude en sus fases iniciales, antes de que alcancen escala (GSMA FASG#33, 2025).

En el contexto español, este requisito de correlación intersectorial tiene una implicación directa sobre la arquitectura de la plataforma de inteligencia propuesta en la Sección 4.9.1: debe ser capaz de ingestar y correlacionar en tiempo real señales heterogéneas procedentes de operadores telco (patrones de SMS, registros de SIM, tráfico anómalo), plataformas digitales (intentos de login fallidos, cambios de credenciales, accesos desde nuevas ubicaciones) y entidades financieras (microtransacciones de prueba, transferencias hacia cuentas sin historial, actividad en horarios atípicos). Ninguno de estos actores, de forma aislada, dispone del cuadro completo. La plataforma de inteligencia es, en este sentido, el instrumento técnico que hace posible la detección preventiva del fraude a escala nacional.

Adicionalmente, el desarrollo de capacidades de detección de contenido sintético y deepfakes —mencionado como Recomendación R5 en la Sección 7.2— constituye una dimensión específica de la detección que requiere inversión diferenciada y colaboración con el ecosistema de investigación en inteligencia artificial. El ciclo de innovación en generación de contenido sintético es actualmente más rápido que el de los sistemas de detección, lo que genera una brecha que solo puede reducirse mediante inversión pública sostenida y colaboración internacional.

Eje 3 — Intervención: interrumpir el fraude en curso con mínima fricción para el usuario legítimo

La intervención actúa sobre campañas de fraude ya detectadas con el objetivo de interrumpirlas antes de que completen el ciclo de monetización. La eficacia de la intervención depende críticamente de tres factores: la velocidad de respuesta (que debe operar a la velocidad del ataque, no del proceso administrativo), la coordinación entre los actores que tienen capacidad de actuación técnica (operadores telco que pueden bloquear números y dominios, plataformas que pueden desactivar cuentas comprometidas, entidades financieras que pueden congelar transacciones sospechosas), y la minimización del daño colateral sobre usuarios legítimos (el bloqueo excesivo o los falsos positivos deterioran la experiencia de usuario y generan desconfianza adicional).

El bloqueo rápido de campañas fraudulentas en el canal telco —especialmente el bloqueo de números emisores de smishing y dominios de phishing— requiere protocolos técnicos y jurídicos pre-acordados entre los operadores y las autoridades, que permitan actuar en minutos en lugar de días. El Reino Unido dispone de un sistema de reporte y bloqueo de smishing con tiempos de respuesta inferiores a dos horas; en España, la ausencia de un

mecanismo equivalente implica que las campañas de smishing documentadas en este trabajo —con decenas o cientos de miles de mensajes enviados en pocas horas— pueden completar su ciclo antes de que se activen las medidas de bloqueo (GSMA FASG#34, 2026).

La desactivación de cuentas comprometidas en plataformas de contenido y la congelación de transacciones sospechosas en el sistema financiero son igualmente dependientes de la velocidad y la coordinación. La interoperabilidad entre los sistemas de gestión de identidad de las plataformas y los sistemas antifraude de las entidades financieras —que permitiría, por ejemplo, que la detección de un account takeover en una plataforma de streaming activara automáticamente un control adicional sobre la cuenta bancaria vinculada— es una capacidad técnica que existe en sistemas propietarios de grandes corporaciones pero que no está disponible como estándar de interoperabilidad para el conjunto del mercado español.

Eje 4 — Persecución: desarticular la infraestructura criminal y recuperar activos

El eje de persecución cierra el ciclo del modelo y actúa sobre las organizaciones criminales que operan el fraude, con el objetivo de desarticular su infraestructura, recuperar los activos defraudados y generar efecto disuasorio sobre futuros actores. La persecución eficaz del fraude digital en plataformas de contenido enfrenta tres desafíos estructurales en el contexto español y europeo.

El primero es la dimensión transnacional de la infraestructura criminal: los actores del fraude operan deliberadamente a través de múltiples jurisdicciones para dificultar la investigación y la recuperación de activos. La inteligencia sobre las organizaciones criminales desarticuladas por el Ministerio del Interior y la Guardia Civil en los últimos años revela que sus estructuras incluyen componentes en Europa del Este, Latinoamérica y el sudeste asiático, con infraestructura técnica alojada en jurisdicciones con baja cooperación judicial (Ministerio del Interior, 2026; Global Initiative Against Transnational Organized Crime, 2026). El refuerzo de los mecanismos de cooperación judicial internacional a través de Europol y Eurojust, y la participación activa de España en iniciativas como la Joint Cybercrime Action Taskforce (J-CAT) de Europol, son condiciones necesarias para que la persecución sea eficaz más allá del nivel doméstico.

El segundo desafío es la trazabilidad financiera de los flujos de blanqueo. La combinación de cuentas mulas, transferencias fraccionadas y conversión a criptomonedas genera una cadena de movimientos que dificulta la recuperación de activos incluso cuando se identifican los responsables. El refuerzo de las capacidades de análisis de blockchain de las Fuerzas y Cuerpos de Seguridad del Estado, la colaboración con los exchanges de criptomonedas regulados para la identificación de flujos sospechosos, y el cumplimiento de las obligaciones de la Directiva AMLD6 en materia de activos digitales son herramientas regulatorias y técnicas disponibles que deben ser utilizadas de forma sistemática (FATF, 2023).

El tercer desafío es la adecuación del marco penal español a la realidad del fraude digital industrializado. El Código Penal tipifica las conductas relevantes —estafa informática, suplantación de identidad, acceso no autorizado a sistemas—, pero la aplicación de circunstancias agravantes vinculadas a la escala, la organización y el uso de infraestructura criminal especializada no siempre se traduce en sentencias que generen suficiente efecto disuasorio. La actualización de la doctrina jurisprudencial sobre fraude

digital en plataformas de contenido, y la formación específica de fiscales y jueces en esta materia, son elementos complementarios e indispensables del eje de persecución (Gobierno de España, 2023; Europol, 2023b).

9.1.1 Síntesis del modelo: interdependencias y priorización

Los cuatro ejes del modelo PDIP no son independientes ni secuenciales en sentido estricto: operan simultáneamente y sus resultados se retroalimentan. Una inversión insuficiente en prevención incrementa el volumen de incidentes que deben ser detectados; una detección tardía limita la eficacia de la intervención; una intervención incompleta protege a los actores criminales frente a la persecución. Sin embargo, dado que los recursos disponibles —tanto en las organizaciones sectoriales como en las administraciones públicas— son finitos, la priorización de inversiones en el ciclo PDIP debe basarse en el análisis del eslabón más débil de la cadena.

En el contexto español actual, ese eslabón es la detección intersectorial en tiempo real y la coordinación operativa en el eje de intervención. Los sistemas de prevención tienen un desarrollo razonable, aunque mejorable en alcance y actualización; los marcos de persecución existen aunque con limitaciones en la cooperación internacional; pero la capacidad de correlacionar señales de fraude entre sectores y de activar bloqueos coordinados en tiempo real es prácticamente inexistente como infraestructura nacional. Esta evaluación orienta la secuencia de implementación del modelo: la creación de la Plataforma Nacional de Inteligencia Antifraude y los protocolos de intervención coordinada entre operadores telco, plataformas y entidades financieras representan la inversión con mayor retorno esperado en la reducción del impacto del fraude digital en España a corto y medio plazo.

Eje	Actores principales / instrumentos clave
Prevención	INCIBE (OSI, formación); operadores telco (campañas); plataformas (security by design, fraud-resistant UX); reguladores (estándares de autenticación)
Detección	INCIBE (PNIA — Plataforma Nacional de Inteligencia Antifraude); operadores telco y plataformas (señales de comportamiento); entidades financieras (alertas de transacciones); UEBA y ML
Intervención	Operadores telco (bloqueo de smishing/dominios); plataformas (desactivación de cuentas comprometidas); entidades financieras (congelación de transacciones); FCSE (operaciones de respuesta rápida)
Persecución	FCSE — Brigada de Investigación Tecnológica, UCO; Europol / Eurojust; fiscalía especializada; exchanges de criptomonedas (trazabilidad AML); reforma del marco penal

Tabla 2 Síntesis del modelo PDIP para la mitigación del fraude digital en plataformas de contenido en España: ejes, actores principales e instrumentos clave. Elaboración propia a partir de GSMA (2025), CFCA (2023), Europol (2023) e INCIBE.

La implementación efectiva del modelo PDIP requiere, en última instancia, un cambio de paradigma en la forma en que los actores del ecosistema digital español conciben su responsabilidad frente al fraude: pasar de una lógica de protección de perímetro propio —cada actor gestiona el fraude que le afecta directamente— a una lógica de

responsabilidad compartida sobre el ecosistema, en la que la seguridad colectiva es un bien público que requiere inversión coordinada, estándares comunes y mecanismos de gobernanza que alineen los incentivos de actores con intereses parcialmente divergentes. El fraude digital ha alcanzado la escala y la sofisticación que convierten esta transición no en una opción, sino en una condición de supervivencia del modelo de economía digital que España aspira a construir.

9.2. Modelo socio-técnico del fraude en plataformas de contenido (MSFPC)

9.2.1 Fundamentos del MSFPC: el fraude como sistema socio-técnico adaptativo

A partir del análisis desarrollado en los capítulos anteriores, este trabajo propone un marco conceptual que integra las distintas dimensiones del fraude en plataformas de contenidos digitales bajo la perspectiva de los sistemas socio-técnicos (Bostrom & Heinen, 1977; Baxter & Sommerville, 2011). En este contexto, el fraude se conceptualiza no como un conjunto de incidentes aislados, sino como un sistema adaptativo, multi-actor y multi-capa, cuya dinámica emerge de la interacción entre componentes tecnológicos, comportamientos humanos, estructuras organizativas y marcos regulatorios.

Este enfoque permite superar las limitaciones de la literatura existente, que ha tendido a abordar el fraude desde perspectivas fragmentadas —técnica, económica o legal— sin capturar las interdependencias entre dichas dimensiones. En particular, el modelo propuesto incorpora explícitamente la coevolución entre atacantes y sistemas de defensa, alineándose con enfoques de sistemas complejos y entornos adversariales en seguridad digital (Anderson, 2020).

La Figura 11 representa la arquitectura conceptual del modelo, estructurada en cuatro capas interrelacionadas: (i) entorno estructural, que define las condiciones del ecosistema digital; (ii) superficie de ataque, donde se materializan las vulnerabilidades asociadas a cuentas, pagos, datos e interfaces; (iii) sistema de fraude, que constituye el núcleo dinámico del modelo; y (iv) sistema de defensa, que agrupa los mecanismos técnicos, organizativos y regulatorios orientados a la mitigación del riesgo.

El elemento central del modelo reside en la conceptualización del fraude como la interacción entre dos dimensiones complementarias: los métodos de ataque, basados en la manipulación cognitiva y conductual del usuario, y los mecanismos técnicos, orientados a la explotación de vulnerabilidades tecnológicas. Esta interacción se formaliza mediante el constructo method–mechanism coupling (MMC), que captura el efecto multiplicativo derivado de combinar vectores sociales y técnicos, y que constituye el principal motor de la efectividad del fraude.

Asimismo, el modelo incorpora un bucle de retroalimentación adaptativa, mediante el cual las mejoras en los sistemas de defensa generan presiones selectivas que incentivan

la evolución de las estrategias de ataque. Este mecanismo introduce una dinámica no lineal en el sistema, en la que el fraude no solo responde a las condiciones del entorno, sino que también se reconfigura en función de las propias medidas de mitigación implementadas.

En conjunto, este marco conceptual permite interpretar el fraude como un fenómeno estructural de la economía digital, caracterizado por su capacidad de adaptación, su lógica económica y su integración en el funcionamiento de las plataformas digitales.

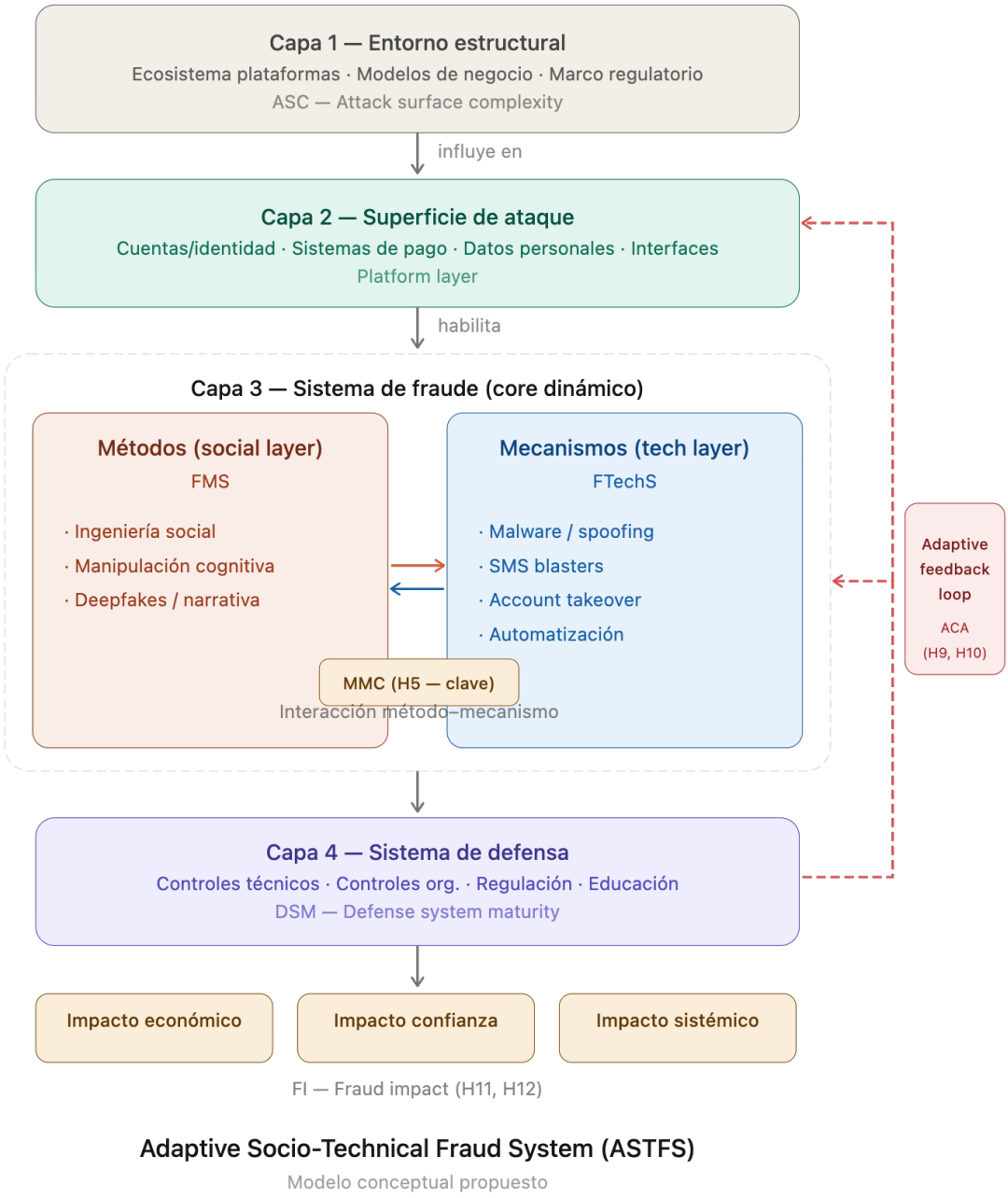

Figura 11 Arquitectura conceptual del Modelo Socio-Técnico del Fraude en Plataformas de Contenido (MSFPC): cuatro capas interrelacionadas (entorno estructural, superficie de ataque,

sistema de fraude, sistema de defensa) con bucle de retroalimentación adaptativa. Elaboración propia.

9.2.2 Desarrollo del modelo y definición de constructos

Con el objetivo de operacionalizar el marco conceptual propuesto y posibilitar su contraste empírico, se desarrolla un modelo causal que especifica las relaciones entre los principales constructos identificados. Este modelo traduce la arquitectura socio-técnica descrita previamente en un conjunto de variables observables y relaciones hipotéticas, facilitando su validación en contextos empíricos.

La Figura 12 presenta el modelo causal resultante, en el que se distinguen tres niveles analíticos: (i) factores estructurales que condicionan la exposición al fraude; (ii) mecanismos generadores del fraude; y (iii) consecuencias del fraude en el ecosistema digital.

En el nivel estructural, se introduce el constructo complejidad de la superficie de ataque (ASC), que refleja el grado de exposición derivado de la arquitectura de la plataforma, incluyendo la diversidad de funcionalidades, la integración de servicios y la heterogeneidad de interfaces. Este constructo actúa como antecedente de la sofisticación del fraude, al ampliar el conjunto de vectores potenciales disponibles para los atacantes.

En el núcleo del modelo se sitúan dos constructos clave: la sofisticación de los métodos de fraude (FMS) y la sofisticación de los mecanismos técnicos (FTechS). El primero captura la capacidad de los atacantes para explotar vulnerabilidades cognitivas y conductuales mediante técnicas de ingeniería social, mientras que el segundo refleja el nivel de complejidad técnica de los vectores de ataque, incluyendo su automatización y escalabilidad.

La interacción entre ambos se conceptualiza mediante el constructo method–mechanism coupling (MMC), modelado como un efecto de moderación que amplifica el impacto conjunto de FMS y FTechS. Este planteamiento permite capturar la naturaleza híbrida del fraude contemporáneo, en la que la combinación de dimensiones sociales y técnicas genera efectos no lineales sobre su efectividad.

Adicionalmente, se incorpora el constructo industrialización del fraude (FIz), que representa el grado en que las actividades fraudulentas adoptan lógicas de producción escalable, incluyendo automatización, especialización funcional y uso de infraestructuras compartidas. Este constructo actúa como mecanismo intermedio que conecta la sofisticación técnica con el impacto del fraude.

En el ámbito de la mitigación, el modelo incluye la madurez del sistema de defensa (DSM), que integra capacidades técnicas, organizativas y regulatorias orientadas a la prevención del fraude. No obstante, en línea con la conceptualización del fraude como sistema adaptativo, se introduce la capacidad adaptativa de los atacantes (ACA) como un

constructo mediador que refleja la habilidad de los atacantes para evolucionar en respuesta a las medidas de defensa. Esta relación captura la dinámica de coevolución atacante–defensor y permite modelar efectos de desplazamiento del fraude en el tiempo.

Finalmente, el modelo incorpora el constructo impacto del fraude (FI) como variable dependiente principal, definido como un fenómeno multidimensional que incluye efectos económicos, deterioro de la confianza del usuario y consecuencias sistémicas en el ecosistema digital. Como extensión, se consideran los efectos del fraude sobre la confianza del usuario (UT) y la sostenibilidad de la plataforma (PS), estableciendo un vínculo entre seguridad, comportamiento del usuario y viabilidad económica de las plataformas digitales.

En conjunto, este modelo permite no solo explicar la generación y evolución del fraude, sino también analizar sus consecuencias y las limitaciones de los sistemas de mitigación en entornos dinámicos y complejos.

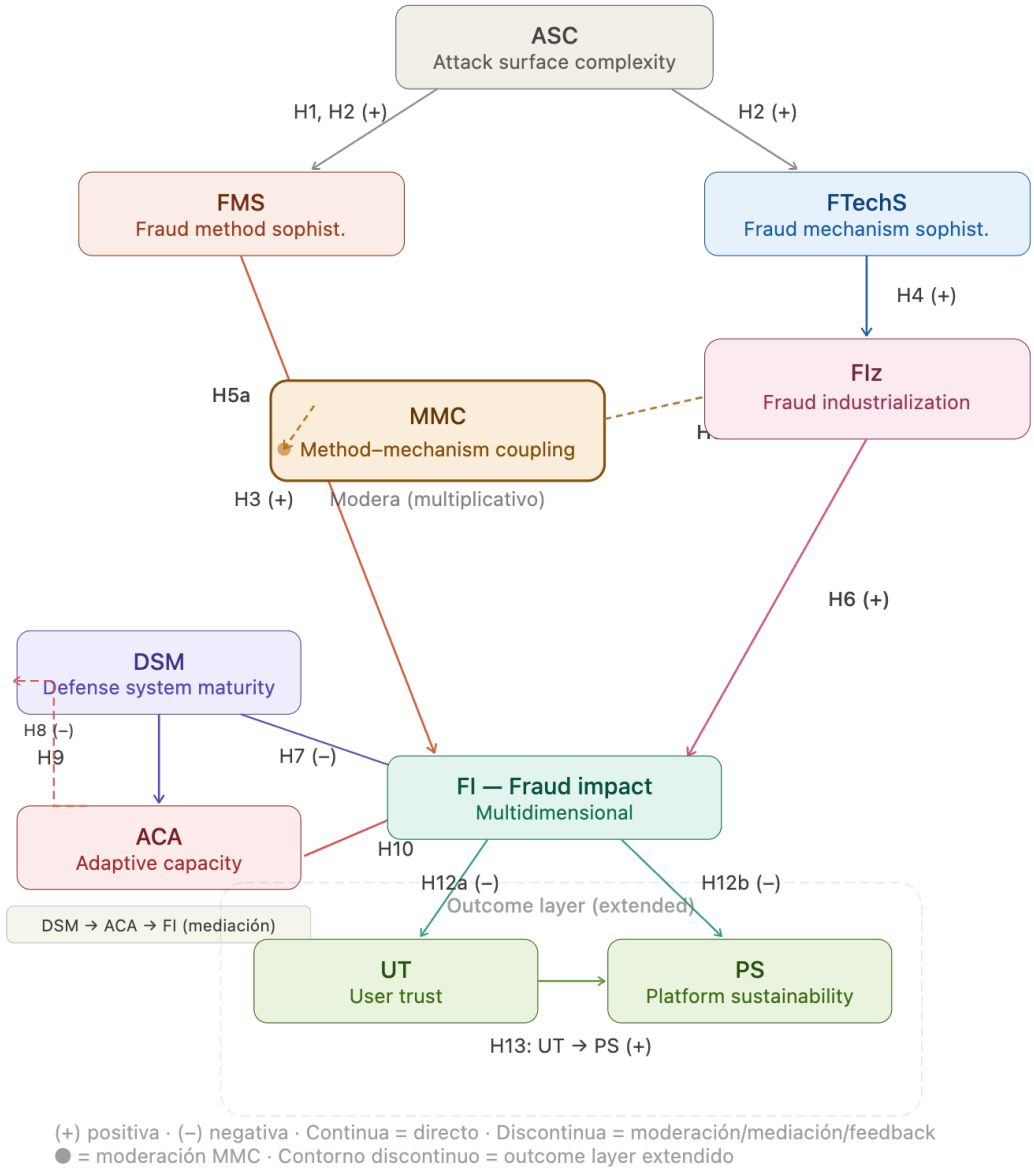

Figura 12 Modelo causal del MSFPC: constructos, relaciones hipotéticas (H1-H13) y tres niveles analíticos (factores estructurales, mecanismos generadores, consecuencias ecosistémicas). Elaboración propia.

9.2.3 Desarrollo de hipótesis

9.2.3.1 Complejidad de la superficie de ataque y sofisticación del fraude

La literatura en sistemas de información y seguridad ha señalado que la complejidad estructural de los sistemas digitales incrementa la exposición a vulnerabilidades y vectores de ataque (Anderson, 2020; ENISA, 2023). En el contexto de plataformas de contenidos digitales, esta complejidad se deriva de la integración de múltiples funcionalidades —cuentas de usuario, sistemas de pago, datos personales e interfaces de interacción— que configuran una superficie de ataque heterogénea y estratificada (Sección 2.2).

Desde una perspectiva socio-técnica, una mayor complejidad de la superficie de ataque (ASC) no solo incrementa el número de puntos potenciales de explotación, sino que también facilita la aparición de vectores híbridos que combinan dimensiones técnicas y sociales. Esto permite a los atacantes desarrollar estrategias más sofisticadas tanto en términos de manipulación del usuario como de explotación tecnológica.

En este sentido, se propone que la complejidad de la superficie de ataque actúa como un factor habilitador de la sofisticación del fraude, tanto en su dimensión metodológica (ingeniería social) como en su dimensión técnica.

H1. La complejidad de la superficie de ataque (ASC) se asocia positivamente con la sofisticación de los métodos de fraude (FMS).

H2. La complejidad de la superficie de ataque (ASC) se asocia positivamente con la sofisticación de los mecanismos de fraude (FTechS).

9.2.3.2 Núcleo del fraude: interacción entre métodos y mecanismos

La investigación en fraude digital ha tendido a analizar de forma separada las dimensiones técnicas (e.g., malware, spoofing) y las dimensiones sociales (e.g., phishing, ingeniería social). Sin embargo, evidencia reciente sugiere que los ataques más efectivos emergen precisamente de la combinación de ambas dimensiones (Levi & Smith, 2021; GSMA, 2025).

La sofisticación de los métodos de fraude (FMS), basada en la manipulación cognitiva y emocional del usuario, incrementa la probabilidad de éxito de los ataques al explotar sesgos conductuales (Sección 4.1). Por su parte, la sofisticación de los mecanismos técnicos (FTechS) permite escalar los ataques, automatizarlos y eludir controles técnicos.

No obstante, el presente trabajo propone que el verdadero motor de la efectividad del fraude reside en la interacción entre ambas dimensiones, conceptualizada como method–mechanism coupling (MMC). Este constructo captura el efecto multiplicativo derivado de combinar ingeniería social y explotación tecnológica, generando ataques más difíciles de detectar y mitigar.

Así, mientras que FMS y FTechS tienen efectos directos sobre el impacto del fraude, su interacción amplifica significativamente dichos efectos.

H3. La sofisticación de los métodos de fraude (FMS) se asocia positivamente con el impacto del fraude (FI).

H4. La sofisticación de los mecanismos de fraude (FTechS) se asocia positivamente con el impacto del fraude (FI).

H5. El acoplamiento método-mecanismo (MMC) modera positivamente la relación entre FMS/FTechS y el impacto del fraude (FI), generando un efecto multiplicativo.

9.2.3.3 Industrialización del fraude

El fraude digital ha evolucionado hacia modelos organizados y escalables, caracterizados por la automatización, la especialización funcional y la existencia de infraestructuras compartidas (Fraud-as-a-Service) (Anderson, 2020; Europol, 2023). Este proceso, conceptualizado como industrialización del fraude (FIz), permite a los atacantes incrementar el volumen de ataques y reducir el coste marginal por operación.

La sofisticación técnica (FTechS) constituye un factor clave en este proceso, ya que habilita la automatización y la escalabilidad de los ataques (Sección 4.2). A su vez, la industrialización del fraude amplifica su impacto al aumentar la frecuencia y alcance de los ataques.

H6. La industrialización del fraude (FIz) se asocia positivamente con el impacto del fraude (FI).

9.2.3.4 Sistema de defensa y dinámica adaptativa

Los sistemas de defensa en plataformas digitales —incluyendo autenticación multifactor, análisis de comportamiento y modelos de inteligencia artificial— tienen como objetivo reducir la incidencia y el impacto del fraude (Sección 6.2). En este sentido, una mayor madurez del sistema de defensa (DSM) debería contribuir a la mitigación del fraude.

H7. La madurez del sistema de defensa (DSM) se asocia negativamente con el impacto del fraude (FI).

Sin embargo, desde la perspectiva de sistemas adaptativos, las medidas de defensa también actúan como mecanismos de presión selectiva que incentivan la evolución de los atacantes (Baxter & Sommerville, 2011). Esto da lugar a una capacidad adaptativa de los atacantes (ACA), que refleja su habilidad para modificar estrategias y eludir controles.

Así, la relación entre defensa y fraude no es estática, sino dinámica: las mejoras en defensa pueden inducir procesos de adaptación en los atacantes, reduciendo la eficacia de dichas defensas en el largo plazo.

H8. La capacidad adaptativa de los atacantes (ACA) reduce la eficacia de los sistemas de defensa (DSM) en la mitigación del fraude.

H9. La madurez del sistema de defensa (DSM) se asocia positivamente con la capacidad adaptativa de los atacantes (ACA).

En este contexto, la capacidad adaptativa actúa como un mecanismo mediador en la relación entre defensa e impacto del fraude.

H10. La capacidad adaptativa de los atacantes (ACA) media la relación entre la madurez del sistema de defensa (DSM) y el impacto del fraude (FI).

9.2.3.5 Impacto del fraude y consecuencias en el ecosistema

El impacto del fraude (FI) se configura como un constructo multidimensional que incluye dimensiones económicas, de confianza y sistémicas (Sección 1.2). Más allá de las

pérdidas directas, el fraude afecta a la percepción de seguridad del usuario y a la sostenibilidad de las plataformas digitales.

Desde la literatura sobre confianza en sistemas digitales (Gillespie, 2018; WEF, 2025), se establece que una mayor incidencia de fraude deteriora la confianza del usuario en las plataformas.

H11. El impacto del fraude (FI) se asocia negativamente con la confianza del usuario (UT).

Asimismo, la confianza del usuario constituye un factor clave para la sostenibilidad de las plataformas digitales, al influir en la adopción, retención y monetización de los servicios.

H12. El impacto del fraude (FI) se asocia negativamente con la sostenibilidad de la plataforma (PS).

Finalmente, la literatura en plataformas digitales sugiere que la confianza del usuario actúa como un determinante directo de la sostenibilidad del ecosistema.

H13. La confianza del usuario (UT) se asocia positivamente con la sostenibilidad de la plataforma (PS).

9.2.4 Operacionalización de constructos y marco para la validación empírica

El modelo MSFPC desarrollado en las secciones precedentes tiene un carácter teórico-propositivo: formaliza relaciones causales entre constructos a partir de la evidencia cualitativa analizada en los Capítulos 1 a 8, pero no ha sido sometido a contraste empírico cuantitativo. Para que las trece hipótesis formuladas sean efectivamente contrastables, es necesario operacionalizar cada constructo mediante indicadores observables, identificar las fuentes de datos potenciales y definir las escalas de medición aplicables.

La tabla siguiente presenta esta operacionalización para los diez constructos del modelo. Cada constructo se descompone en indicadores propuestos que reflejan las dimensiones conceptuales descritas en la Sección 9.2.2, se vincula a fuentes de datos accesibles en el contexto español y europeo, y se asocia a un tipo de escala coherente con la naturaleza del indicador. Esta operacionalización no constituye un diseño de investigación cerrado, sino un marco de referencia que orienta la validación empírica futura del modelo.

Constructo	Indicadores propuestos	Fuente de datos potencial	Escala
ASC — Complejidad de la superficie de ataque	N.º de funcionalidades que gestionan datos sensibles (cuentas, pagos, publicidad, datos personales); n.º de interfaces externas (APIs, pasarelas); diversidad de métodos de autenticación; volumen de datos personales por usuario activo	Análisis técnico de arquitectura; informes de transparencia DSA; documentación de APIs públicas	Ratio (índice compuesto normalizado 0–1)
FMS — Sofisticación de los métodos de fraude	N.º de sesgos cognitivos explotados por campaña (urgencia, autoridad, escasez, reciprocidad; Cialdini, 2009); grado de personalización (genérico / segmentado /	INCIBE-CERT (informes de incidentes); Ministerio del Interior (denuncias); GSMA FASG; Europol IOCTA	Ordinal 3 niveles por dimensión, agregable en índice compuesto

Constructo	Indicadores propuestos	Fuente de datos potencial	Escala
	individualizado); n.º de canales simultáneos; duración del ciclo de enganche		
FTechS — Sofisticación técnica del fraude	Grado de automatización (manual / semi / completo); n.º de capas defensivas superadas; sofisticación del canal (SMS → Sender ID suplantado → SMS blaster); uso de tecnologías avanzadas (deepfake, reverse proxy, malware cifrado)	Análisis forenses post-incidente; CDR de operadores; informes técnicos GSMA FASG#33/#34	Ordinal, derivada de taxonomía del Capítulo 4
MMC — Acoplamiento método-mecanismo	Producto de puntuaciones estandarizadas de FMS × FTechS; alternativamente: categórica (ataque solo social / solo técnico / híbrido)	Clasificación de incidentes INCIBE-CERT y Europol IOCTA	Interacción en SEM (producto FMS × FTechS) o categórica 3 niveles
FIz — Industrialización del fraude	Volumen de ataques por campaña; existencia de infraestructura compartida (SIM farms, paneles C2); grado de especialización funcional; coste marginal estimado por víctima	GSMA FASG (volúmenes, SIM boxes); Europol (desarticulaciones); precios documentados en mercados FaaS	Ratio (volumen, coste) + ordinal (especialización)
DSM — Madurez del sistema de defensa	Nivel de madurez GSMA Fraud Management Framework (5 niveles); tasa de adopción de MFA; MTTD; MTTR; existencia de protocolos de coordinación intersectorial	Evaluaciones GSMA; informes TIBER-EU (BCE); datos operativos CSIRT nacionales	Ordinal (madurez GSMA) + ratio (MTTD, MTTR)
ACA — Capacidad adaptativa de los atacantes	Tiempo medio entre implementación defensiva y aparición de vector de evasión; tasa de desplazamiento entre canales tras intervención; diversificación de vectores por organización criminal	Informes longitudinales GSMA FASG; estadísticas de desplazamiento de reguladores (ComReg, 2025); tendencias Europol IOCTA	Ratio (tiempos) + ordinal (diversificación)
FI — Impacto del fraude	Pérdidas económicas directas (€); n.º de incidentes reportados; tasa de churn atribuible a fraude; variación en adopción de servicios post-incidente	Banco de España; Visa; GSMA Global Fraud Loss Survey; INCIBE; Ministerio del Interior; datos sectoriales	Ratio (pérdidas, incidentes) + ordinal (percepción)
UT — Confianza del usuario	Percepción de seguridad; disposición a compartir datos de pago; disposición a mantener suscripción post-incidente; confianza en mecanismos de protección	Encuesta a usuarios (escala Likert 7 puntos, adaptando Gillespie, 2018); ONTSI; Eurobarómetro; encuestas INCIBE-OSI	Intervalo (Likert 7 puntos)
PS — Sostenibilidad de la plataforma	Tasa de retención de suscriptores; ARPU; coste antifraude como % del ingreso; posición competitiva relativa	Informes financieros públicos de plataformas; análisis sectoriales Omdia; datos agregados GSMA y CFCA	Ratio

Tabla 3 Operacionalización de constructos del modelo MSFPC: indicadores propuestos, fuentes de datos y escalas de medición. Elaboración propia.

Constructos de nivel estructural

ASC — Complejidad de la superficie de ataque. Este constructo refleja el grado de exposición derivado de la arquitectura de la plataforma. Los indicadores propuestos incluyen: número de funcionalidades integradas que gestionan datos sensibles (cuentas, pagos, publicidad, datos personales), número de interfaces externas (APIs, pasarelas de pago, integraciones con terceros), diversidad de métodos de autenticación soportados, y volumen de datos personales procesados por usuario activo. La fuente de datos primaria sería un análisis técnico de la arquitectura de la plataforma, complementado con datos

publicados en informes de transparencia (DSA) y documentación de APIs públicas. La escala es una escala compuesta de ratio, construida como índice ponderado de las dimensiones anteriores, normalizable entre 0 y 1 para comparaciones entre plataformas.

Constructos del núcleo del fraude

FMS — Sofisticación de los métodos de fraude. Captura la complejidad de las técnicas de ingeniería social empleadas. Los indicadores propuestos son: número de sesgos cognitivos explotados por campaña (urgencia, autoridad, escasez, reciprocidad, según la taxonomía de Cialdini, 2009), grado de personalización del ataque (genérico, segmentado, individualizado), número de canales utilizados simultáneamente (SMS, email, redes sociales, llamada telefónica), y duración del ciclo de enganche (desde el primer contacto hasta la acción fraudulenta). Las fuentes de datos incluyen los informes de incidentes de INCIBE-CERT, las bases de datos de denuncias del Ministerio del Interior, y los análisis de campañas de phishing documentados por GSMA FASG y Europol. La escala es ordinal de tres niveles (bajo, medio, alto) para cada dimensión, agregable en un índice compuesto.

FTechS — Sofisticación técnica del fraude. Refleja el nivel de complejidad tecnológica de los vectores de ataque. Los indicadores propuestos incluyen: uso de automatización (manual, semiautomático, completamente automatizado), capacidad de evasión de controles (número de capas de defensa superadas), sofisticación del canal de distribución (SMS genérico, SMS con Sender ID suplantado, SMS blaster fuera de red), y uso de tecnologías avanzadas (deepfake, reverse proxy, malware con exfiltración cifrada). Las fuentes de datos son los análisis forenses post-incidente, los CDR de operadores telco, y los informes técnicos de GSMA FASG#33 y FASG#34. La escala es ordinal, con niveles derivados de la taxonomía de vectores del Capítulo 4.

MMC — Acoplamiento método-mecanismo. Constructo de interacción que captura el efecto multiplicativo de combinar FMS y FTechS. Se operacionaliza como el producto de las puntuaciones estandarizadas de FMS y FTechS para cada incidente o campaña analizada. Alternativamente, puede medirse como una variable categórica de tres niveles: ataque puramente social (MMC bajo), ataque puramente técnico (MMC bajo), ataque híbrido (MMC alto). La fuente de datos es la clasificación de incidentes en las bases de INCIBE-CERT y Europol IOCTA, que categorizan los vectores utilizados en cada caso. En un modelo de ecuaciones estructurales (SEM), el MMC se modelaría como un término de interacción entre FMS y FTechS, no como un constructo latente independiente.

Constructo de escalabilidad

FIz — Industrialización del fraude. Captura el grado en que la actividad fraudulenta adopta lógicas de producción a escala. Los indicadores propuestos son: volumen de ataques por campaña (número de mensajes/intentos), existencia de infraestructura compartida (SIM farms, dominios reutilizados, paneles C2 como servicio), grado de especialización funcional (separación entre proveedor de infraestructura, operador de campaña y red de monetización), y coste marginal estimado por víctima captada. Las fuentes de datos incluyen los datos operativos de GSMA FASG (volúmenes de SMS fraudulentos, SIM boxes detectadas), los informes de Europol sobre desarticulación de redes, y los precios documentados en mercados de Fraud-as-a-Service. La escala es de

ratio para los indicadores cuantitativos (volumen, coste) y ordinal para los cualitativos (especialización).

Constructos de defensa y adaptación

DSM — Madurez del sistema de defensa. Integra las capacidades técnicas, organizativas y regulatorias de mitigación. Se propone operacionalizar este constructo mediante el nivel de madurez del GSMA Fraud Management Framework (2025), que ya define una escala progresiva de cinco niveles desde la defensa perimetral aislada hasta la inteligencia compartida en tiempo real. Los indicadores complementarios incluyen: tasa de adopción de MFA entre usuarios, tiempo medio de detección de incidentes (MTTD), tiempo medio de respuesta (MTTR), y existencia de protocolos de coordinación intersectorial. Las fuentes de datos son las evaluaciones de madurez del GSMA, los informes TIBER-EU del BCE, y los datos operativos de los CSIRT nacionales. La escala es ordinal (niveles de madurez GSMA) combinada con métricas de ratio (MTTD, MTTR).

ACA — Capacidad adaptativa de los atacantes. Refleja la habilidad de los atacantes para modificar sus estrategias en respuesta a las defensas. Los indicadores propuestos son: tiempo medio entre la implementación de una medida defensiva y la aparición de un vector de evasión documentado, tasa de desplazamiento del fraude entre canales tras una intervención regulatoria (por ejemplo, de SMS a P2P tras la implementación de registros de Sender ID, como documentó ComReg, 2025), y diversificación de vectores de ataque por organización criminal. Las fuentes de datos son los informes longitudinales de GSMA FASG, las estadísticas de desplazamiento documentadas por reguladores nacionales, y los análisis de tendencias de Europol IOCTA. La escala es de ratio para los indicadores temporales y ordinal para la diversificación.

Constructos de impacto y consecuencias

FI — Impacto del fraude. Constructo multidimensional que integra tres subdimensiones: económica (pérdidas directas en euros), de confianza (deterioro de la percepción de seguridad) y sistémica (efectos sobre la adopción de servicios digitales). Los indicadores propuestos son: volumen de pérdidas económicas por fraude (datos de Banco de España, Visa, GSMA Global Fraud Loss Survey), número de incidentes reportados (INCIBE, Ministerio del Interior), tasa de churn atribuible a fraude (datos sectoriales de operadores y plataformas), y variación en la adopción de servicios digitales post-incidente. La escala combina ratio (pérdidas, incidentes) con ordinal (percepción de seguridad, medida mediante encuesta).

UT — Confianza del usuario. Se operacionaliza mediante escalas validadas en la literatura de sistemas de información, adaptadas al contexto de plataformas de contenido. Los indicadores propuestos incluyen: percepción de seguridad de la plataforma, disposición a compartir datos de pago, disposición a mantener la suscripción tras un incidente de fraude, y nivel de confianza en los mecanismos de protección de la plataforma. La fuente de datos primaria sería una encuesta a usuarios basada en escalas Likert de 7 puntos, adaptando instrumentos validados como el de Gillespie (2018) para confianza en plataformas digitales. Fuentes secundarias incluyen los datos de encuestas publicadas por ONTSI, Eurobarómetro y las encuestas de percepción de seguridad de INCIBE-OSI.

PS — Sostenibilidad de la plataforma. Captura la viabilidad económica y competitiva de la plataforma en presencia de fraude. Los indicadores propuestos son: tasa de retención de suscriptores, ingresos medios por usuario (ARPU), coste operativo de la función antifraude como porcentaje del ingreso, y posición competitiva relativa en el mercado. Las fuentes de datos son los informes financieros públicos de las plataformas, los análisis sectoriales de Omdia, y los datos agregados de la industria publicados por GSMA y CFCA. La escala es de ratio.

Enfoque de validación propuesto

La validación empírica del modelo MSFPC puede abordarse mediante tres estrategias complementarias, de menor a mayor complejidad.

La primera estrategia, de viabilidad inmediata, consiste en un estudio Delphi con panel de expertos sectoriales (profesionales de auditoría de fraude, responsables de seguridad de operadores telco y plataformas, analistas de INCIBE y GSMA FASG) para validar la relevancia de los constructos, la plausibilidad de las hipótesis y la adecuación de los indicadores propuestos. Este enfoque es ejecutable a corto plazo con los recursos del grupo de investigación y los contactos institucionales del proyecto PECIEE.

La segunda estrategia, de complejidad intermedia, consiste en un análisis cuantitativo basado en datos secundarios agregados (informes GSMA, Europol, Banco de España, INCIBE) para estimar las relaciones entre constructos mediante técnicas de correlación y regresión. Este enfoque permite una primera aproximación cuantitativa sin requerir datos primarios, aunque con las limitaciones inherentes a la heterogeneidad de las fuentes.

La tercera estrategia, de mayor rigor metodológico, consiste en la estimación del modelo causal completo mediante modelización de ecuaciones estructurales con mínimos cuadrados parciales (PLS-SEM), utilizando datos primarios de encuestas a expertos y datos operativos de plataformas. Este enfoque permitiría estimar simultáneamente las relaciones directas (H1-H4, H6-H7, H11-H13), los efectos de moderación (H5) y los efectos de mediación (H8-H10), proporcionando la validación más completa del modelo. La selección de PLS-SEM sobre CB-SEM se justifica por la naturaleza exploratoria del modelo y el tamaño de muestra previsiblemente limitado en la fase inicial de validación (Hair et al., 2017).

10. Discusión

Este capítulo integra los hallazgos de los capítulos anteriores en una discusión que conceptualiza el fraude como un sistema socio-técnico adaptativo, e incluye una evaluación crítica de los enfoques actuales.

10.1 Interpretación de los hallazgos: el fraude como sistema socio-técnico adaptativo

El análisis desarrollado en los Capítulos 1 a 8 converge en una conclusión central: el fraude en plataformas de contenidos digitales no es un conjunto de incidentes aislados susceptibles de control mediante intervenciones técnicas puntuales, sino un sistema socio-técnico adaptativo cuya dinámica emerge de la interacción entre tecnología, comportamiento humano, modelos de negocio y marcos regulatorios.

Esta conclusión se sustenta en cuatro líneas de evidencia transversales. La primera es la naturaleza multi-capa del fenómeno: el Capítulo 2 muestra que las cinco superficies de ataque identificadas (cuentas, suscripciones, pagos, publicidad, datos personales) no operan de forma independiente, sino que un compromiso en una capa habilita la explotación de las restantes. Los tres casos end-to-end del Capítulo 8 ilustran esta cascada operativamente: en el Caso A, el compromiso de credenciales (capa 1) alimenta la explotación de la suscripción (capa 2) y del sistema de pago (capa 3), con la monetización final operando en la capa de datos personales (capa 5) a través de las cuentas mula. Esta interdependencia entre capas es precisamente lo que los modelos de fraude tradicionales, centrados en tipologías discretas, no capturan.

La segunda línea de evidencia es la convergencia método-mecanismo. El Capítulo 4 documenta que los ataques más eficaces no son puramente técnicos (malware, exploits) ni puramente sociales (phishing genérico), sino híbridos que combinan manipulación cognitiva con explotación tecnológica. El constructo *method-mechanism coupling* (MMC) del MSFPC formaliza esta observación empírica recurrente: el reverse proxy phishing del Caso A, el deepfake con pig butchering del Caso B, y el SpyLoan con extorsión reputacional del Caso C son manifestaciones de este acoplamiento en tres contextos distintos.

La tercera línea es la dinámica adaptativa atacante-defensor. La evidencia del Capítulo 6 muestra que cada avance defensivo (MFA, UEBA, WAF) induce una respuesta adaptativa de los atacantes documentada empíricamente: el MFA obligatorio desplaza los ataques hacia SIM swapping y reverse proxy; los filtros de SMS desplazan el fraude hacia canales P2P; los registros de Sender ID desplazan los vectores hacia short codes (ComReg, 2025). Esta coevolución, formalizada en el MSFPC mediante la capacidad adaptativa de los atacantes (ACA) como mediador entre defensa e impacto, trasciende la lógica lineal de "más defensa = menos fraude" que subyace a los marcos operativos convencionales.

La cuarta línea es la industrialización económica del fraude. La infraestructura Fraud-as-a-Service documentada en la Sección 4.7, con su desacoplamiento de fases, especialización funcional y economías de escala, opera con una lógica de mercado que

responde a incentivos económicos análogos a los de cualquier industria de servicios. El constructo industrialización del fraude (FIz) del MSFPC captura esta dimensión, conectando la sofisticación técnica con el impacto sistémico a través de la reducción del coste marginal por operación fraudulenta.

10.2 Posicionamiento del MSFPC respecto a modelos existentes

La conceptualización del fraude como fenómeno susceptible de modelización formal tiene una trayectoria consolidada en la literatura, aunque predominantemente centrada en el fraude financiero y corporativo. El posicionamiento del MSFPC requiere una comparación sistemática con los marcos teóricos previos para delimitar su contribución específica.

El modelo más influyente en la literatura clásica es el Triángulo del Fraude (Cressey, 1953), que identifica tres condiciones necesarias para la ocurrencia de fraude: presión percibida, oportunidad y racionalización. El Diamante del Fraude (Wolfe y Hermanson, 2004) añade una cuarta dimensión —la capacidad del perpetrador—, reconociendo que la oportunidad por sí sola no basta sin la competencia técnica para explotarla. Ambos modelos han sido extraordinariamente productivos en el ámbito del fraude corporativo e interno, pero presentan limitaciones significativas cuando se aplican al fraude digital en plataformas: se centran en el perpetrador individual y en sus motivaciones psicológicas, sin abordar la dimensión sistémica del fraude organizado, la interacción entre atacante y víctima como proceso dinámico, ni las características estructurales del ecosistema digital que configuran la superficie de ataque. El MSFPC desplaza el foco desde el perpetrador individual hacia el sistema en su conjunto, integrando la estructura de la plataforma (ASC), las capacidades del atacante (FMS, FTechS) y la respuesta del ecosistema (DSM, ACA) en un modelo causal unificado.

La Teoría de las Actividades Rutinarias (Cohen y Felson, 1979), ampliamente utilizada en criminología, propone que el delito requiere la convergencia en tiempo y espacio de un delincuente motivado, un objetivo adecuado y la ausencia de un guardián capaz. Este marco ha sido adaptado al cibercrimen por diversos autores, pero su aplicación al fraude en plataformas digitales enfrenta una limitación estructural: en el entorno digital, la convergencia espacio-temporal es permanente (el atacante siempre puede alcanzar al objetivo) y el concepto de "guardián" se fragmenta entre múltiples actores (operador telco, plataforma, entidad financiera) que no comparten información. El MSFPC aborda esta fragmentación explícitamente mediante la madurez del sistema de defensa (DSM) como constructo agregado que integra las capacidades de todos los actores, y mediante las hipótesis H7-H10, que modelan la relación entre defensa, adaptación del atacante e impacto como un bucle dinámico, no como una condición estática de presencia/ausencia de guardián.

El enfoque de economía de la seguridad (Anderson, 2020) aporta una perspectiva complementaria al analizar el fraude desde la lógica de incentivos económicos, costes de protección y externalidades. Este enfoque fundamenta el constructo de industrialización (FIz) del MSFPC, pero Anderson se centra en el análisis de costes agregados sin proponer un modelo causal que relacione constructos específicos ni formular hipótesis

contrastables. El MSFPC extiende esta perspectiva al operacionalizar la dimensión económica como un constructo integrado en una red de relaciones causales.

El GSMA Fraud Management Framework (GSMA, 2025) es el marco operativo más relevante para el sector de telecomunicaciones, con un modelo de madurez progresivo que va desde la defensa perimetral aislada hasta la inteligencia compartida en tiempo real. El modelo PDIP propuesto en la Sección 9.1 de este trabajo adopta y extiende este marco para el contexto español. Sin embargo, el GSMA Framework es un modelo de madurez operativo, no un modelo causal explicativo: describe niveles de capacidad pero no formaliza las relaciones entre variables ni permite derivar hipótesis contrastables sobre por qué determinadas configuraciones defensivas son más eficaces que otras. El MSFPC complementa al GSMA Framework al proporcionar la estructura teórica que explica las dinámicas que el modelo operativo gestiona.

Finalmente, la perspectiva de sistemas socio-técnicos (Bostrom y Heinen, 1977; Baxter y Sommerville, 2011), que constituye el fundamento epistemológico del MSFPC, aporta el marco general de interacción entre componentes técnicos y sociales, pero no ha sido aplicada previamente de forma específica al fraude en plataformas de contenido digital ni ha generado un modelo causal operacionalizado con constructos e hipótesis para este dominio.

En síntesis, la contribución diferencial del MSFPC respecto a los modelos existentes reside en tres elementos: primero, la integración de las dimensiones social (FMS) y técnica (FTechS) mediante un constructo de interacción (MMC) que captura el efecto multiplicativo de la convergencia; segundo, la modelización explícita de la dinámica adaptativa atacante-defensor mediante la mediación de ACA, que introduce no-linealidad en la relación entre defensa e impacto; y tercero, la extensión del modelo hasta las consecuencias ecosistémicas del fraude (confianza del usuario, sostenibilidad de la plataforma), cerrando la cadena causal desde los factores estructurales hasta el impacto sobre la viabilidad del ecosistema digital.

10.3 Evaluación crítica de los enfoques actuales de mitigación

El análisis transversal de los capítulos evidencia cuatro limitaciones estructurales en los enfoques vigentes de mitigación del fraude digital, cada una de las cuales refuerza la necesidad del enfoque sistémico que este trabajo propone.

La primera limitación es la fragmentación del enfoque de seguridad. Los controles técnicos documentados en el Capítulo 6 operan dentro del perímetro de cada organización individual, sin capacidad de correlación intersectorial. Esta fragmentación no es una deficiencia operativa subsanable con más tecnología, sino un problema de arquitectura institucional: los actores del ecosistema (operadores telco, plataformas, entidades financieras) gestionan el fraude como un riesgo propio en lugar de como un riesgo compartido. El Caso A del Capítulo 8 ilustra esta limitación con precisión: las tres señales necesarias para interrumpir la cadena de ataque estuvieron disponibles en tres organizaciones distintas, pero ningún mecanismo permitió correlacionarlas en tiempo real. La propuesta de la PNIA (Recomendación R1) aborda directamente esta carencia.

La segunda limitación es el predominio de modelos reactivos. Los sistemas de detección descritos en las Secciones 6.2.2 a 6.2.4 operan mayoritariamente sobre eventos ya producidos: transacciones ejecutadas, accesos completados, contenido publicado. Esta lógica post-evento es estructuralmente inadecuada frente a un fraude que opera en tiempo real y que completa su ciclo de monetización en horas (Caso A) o que se ejecuta mediante transferencias voluntarias del propio usuario (Caso B). La transición hacia modelos predictivos y preventivos —security by design, fraud-resistant UX, autenticación adaptativa— que las Recomendaciones R6 y R9 proponen requiere no solo inversión tecnológica, sino un cambio en la concepción del fraude: de incidente a gestionar hacia riesgo de diseño a prevenir.

La tercera limitación es la asimetría adaptativa entre atacantes y defensores. La evidencia del Capítulo 4 muestra que las organizaciones criminales innovan con mayor velocidad que los sistemas defensivos: la adopción de deepfakes para fraude de inversión, el despliegue de SMS blasters fuera del plano de control de red, y la externalización de servicios de ataque mediante modelos Fraud-as-a-Service representan innovaciones ofensivas que han precedido sistemáticamente a las respuestas defensivas correspondientes. Esta asimetría se explica en parte por la diferencia en los ciclos de decisión: los atacantes operan sin restricciones regulatorias, contractuales ni de gobernanza corporativa, mientras que los defensores deben cumplir procesos de aprobación, normativa de protección de datos y estándares de interoperabilidad. El constructo ACA del MSFPC formaliza esta dinámica y las hipótesis H8-H10 permiten modelar sus efectos sobre la eficacia de las defensas.

La cuarta limitación es la desalineación regulatoria. El Capítulo 5 documenta un marco normativo que distribuye responsabilidades de forma asimétrica entre actores y que no refleja la naturaleza distribuida del fraude moderno. La PSD2 concentra la responsabilidad en los proveedores de servicios de pago; la DSA se centra en contenido ilegal y desinformación, con atención limitada al fraude como vector específico; la NIS2 refuerza la resiliencia de infraestructuras críticas pero no incluye mecanismos de coordinación operativa entre sectores frente al fraude. La evolución hacia PSD3 y los desarrollos regulatorios en jurisdicciones como Singapur y el Reino Unido (Ramsey, 2024) apuntan hacia modelos de responsabilidad compartida, pero su implementación efectiva requiere la infraestructura de inteligencia intersectorial que el modelo PDIP y las diez recomendaciones de este trabajo proponen.

11. Conclusiones

Este capítulo presenta las conclusiones del estudio, incluyendo la síntesis de las contribuciones teóricas, empíricas y prácticas, las implicaciones del trabajo, sus limitaciones y las líneas de investigación futura.

11.1 Síntesis y contribución principal

Este trabajo ha analizado el fraude en plataformas de contenidos digitales desde una perspectiva integradora, demostrando que no puede entenderse como un conjunto de incidentes aislados, sino como un fenómeno estructural de la economía digital contemporánea.

La principal contribución del estudio reside en la propuesta de un modelo conceptual que interpreta el fraude como un sistema socio-técnico adaptativo, caracterizado por la interacción entre dimensiones tecnológicas, conductuales y organizativas, así como por la existencia de dinámicas de coevolución entre atacantes y sistemas de defensa.

En este contexto, el trabajo introduce el constructo method–mechanism coupling (MMC) como mecanismo central para explicar la efectividad del fraude, y formaliza la dinámica adaptativa atacante–defensor mediante la incorporación de la capacidad adaptativa de los atacantes (ACA). Asimismo, se integra la dimensión económica del fraude a través del concepto de industrialización (FIz), conectando la sofisticación técnica con el impacto sistémico del fenómeno.

11.2. Contribución del estudio

11.2.1 Contribución teórica

La contribución teórica principal del estudio es el Modelo Socio-Técnico del Fraude en Plataformas de Contenido (MSFPC), que integra cuatro dimensiones previamente abordadas de forma fragmentada en la literatura: la dimensión estructural del ecosistema digital, que define la superficie de ataque como un constructo de cinco capas (Capítulo 2); la dimensión conductual de la ingeniería social, formalizada mediante el constructo FMS (Capítulo 4); la dimensión técnica del fraude, capturada por FTechS (Capítulo 4); y la dimensión institucional y regulatoria (Capítulo 5).

Dentro de este marco, el estudio introduce tres aportaciones específicas a la teoría. Primera, el constructo *method–mechanism coupling* (MMC), que formaliza el efecto multiplicativo de la convergencia entre ingeniería social y explotación tecnológica — un fenómeno documentado empíricamente en los Capítulos 4 y 8 pero no modelizado previamente en la literatura. Segunda, la incorporación de la capacidad adaptativa de los atacantes (ACA) como mediador entre la madurez defensiva y el impacto del fraude, lo que introduce una no-linealidad en la relación defensa-impacto que los modelos estáticos existentes (Triángulo del Fraude, Teoría de las Actividades Rutinarias) no capturan, tal como se discute en la Sección 10.2. Tercera, la extensión del modelo hasta las consecuencias ecosistémicas — confianza del usuario (UT) y sostenibilidad de la

plataforma (PS) —, cerrando la cadena causal desde los factores estructurales hasta la viabilidad económica del ecosistema digital.

El conjunto de trece hipótesis (H1-H13) y la tabla de operacionalización de constructos (Tabla 3) proporcionan una base formal para la validación empírica del modelo, como se detalla en la Sección 9.2.4.

11.2.2 Contribución empírica y de síntesis

El estudio realiza una integración sistemática de fuentes heterogéneas — literatura académica, informes sectoriales (GSMA FASG#33/#34, CFCA, Europol IOCTA), datos institucionales (Banco de España, Ministerio del Interior, INCIBE) y evidencia empírica reciente — que permite construir un panorama unificado del fenómeno a través de seis dimensiones: tipologías de fraude (Capítulo 3), vectores técnicos y sociales (Capítulo 4), impactos económicos y sociales (Sección 1.2), marco regulatorio (Capítulo 5), mecanismos de defensa (Capítulo 6) y evidencia operativa (Capítulo 8).

Los tres casos end-to-end del Capítulo 8, construidos como tipificaciones compuestas a partir de patrones documentados en múltiples fuentes primarias, aportan una contribución empírica específica al ilustrar cómo las cadenas de ataque multi-vector explotan las discontinuidades entre actores — una dinámica que la literatura ha descrito de forma genérica pero que estos casos concretan operativamente en el contexto español.

11.2.3 Contribución práctica

El estudio genera tres productos orientados a la aplicación directa. El modelo PDIP (Sección 9.1) proporciona un marco de actuación de cuatro ejes con actores asignados y puntos de intervención derivados del análisis de vectores, adaptado al contexto institucional español. Las diez recomendaciones estratégicas del Capítulo 7, especialmente la propuesta de creación de la Plataforma Nacional de Inteligencia Antifraude (PNIA, Recomendación R1), la propuesta de un sello de confianza antifraude (Recomendación R9) y el marco de indicadores PDIP (Tabla 1), constituyen instrumentos operativos diseñados para su implementación por INCIBE y los actores del ecosistema. El modelo INCIBE como orquestador del sistema antifraude (Sección 7.6) articula las tres dimensiones funcionales — inteligencia, coordinación y resiliencia — que configuran el rol institucional propuesto.

11.3. Implicaciones teóricas

Desde una perspectiva teórica, este estudio contribuye a la literatura en sistemas de información y seguridad digital al extender los enfoques socio-técnicos tradicionales hacia una conceptualización dinámica y adaptativa del fraude.

En particular, el modelo propuesto:

- **Integra dimensiones previamente analizadas** de forma aislada (técnica, conductual, económica y regulatoria).

- **Introduce el concepto de acoplamiento método-mecanismo** como un constructo explicativo de la naturaleza híbrida del fraude.

- **Incorpora dinámicas de coevolución entre atacantes y defensores**, alineadas con teorías de sistemas complejos y entornos adversariales.

Estas aportaciones permiten avanzar hacia una comprensión más completa del fraude como fenómeno sistémico, superando enfoques estáticos centrados en tipologías o controles individuales.

11.4. Implicaciones prácticas

Desde el punto de vista aplicado, los resultados del estudio tienen implicaciones relevantes para el diseño y gestión de sistemas antifraude en plataformas digitales.

En primer lugar, evidencian la necesidad de adoptar enfoques multi-capa y holísticos, que integren controles técnicos, organizativos y regulatorios, en lugar de soluciones aisladas.

En segundo lugar, subrayan la importancia de incorporar la seguridad en el diseño de las plataformas (security by design), prestando especial atención a la reducción de la superficie de ataque y a la protección de la identidad digital del usuario.

En tercer lugar, el reconocimiento de la naturaleza adaptativa del fraude implica que los sistemas de defensa deben evolucionar hacia modelos proactivos y dinámicos, capaces de anticipar la evolución de los ataques y no solo reaccionar a ellos.

Finalmente, el estudio destaca la necesidad de fortalecer la cooperación entre actores del ecosistema —plataformas, operadores, entidades financieras y reguladores— para abordar el fraude como un problema sistémico.

11.5. Limitaciones del estudio

Este trabajo presenta varias limitaciones que deben ser consideradas en la interpretación de los resultados y en la evaluación del alcance de sus conclusiones.

En primer lugar, el modelo MSFPC tiene un carácter teórico-propositivo. Las trece hipótesis formuladas están fundamentadas en evidencia cualitativa y en la convergencia de fuentes sectoriales, institucionales y académicas, pero no han sido sometidas a contraste empírico cuantitativo. La Sección 9.2.4 propone un marco de operacionalización de constructos y tres estrategias de validación (estudio Delphi, análisis de datos secundarios, modelización PLS-SEM), pero su ejecución queda fuera del alcance de este trabajo. Hasta que las hipótesis sean contrastadas con datos primarios, la capacidad predictiva del modelo debe interpretarse con cautela.

En segundo lugar, la naturaleza agregada de algunos constructos — particularmente la capacidad adaptativa de los atacantes (ACA) y la industrialización del fraude (FIz) — puede ocultar dinámicas específicas que varían significativamente en función del tipo de plataforma, la geografía y el entorno regulatorio. Un operador de streaming en España enfrenta una superficie de ataque diferente a la de un marketplace de apps en el sudeste asiático, y el modelo no captura estas diferencias contextuales en su formulación actual.

En tercer lugar, una proporción significativa de la evidencia empírica utilizada procede de fuentes sectoriales y corporativas — GSMA, Kaspersky, McAfee, BioCatch, ACI Worldwide — que, aunque aportan datos operativos de alto valor no disponibles en la literatura académica convencional, no son fuentes independientes: sus informes pueden reflejar sesgos asociados a sus intereses comerciales, la promoción de soluciones propias o la selección de datos que favorecen narrativas específicas. El estudio mitiga esta limitación mediante triangulación (contraste con fuentes institucionales como Europol, Banco de España e INCIBE, y con literatura académica revisada por pares), pero el riesgo de sesgo no puede eliminarse completamente.

En cuarto lugar, los datos de impacto económico citados a lo largo del documento proceden de fuentes que emplean metodologías no siempre comparables entre sí. Las cifras del GSMA Global Fraud Loss Survey, los informes de Visa España, las estimaciones del World Economic Forum y los datos de la Global Initiative Against Transnational Organized Crime difieren en sus definiciones de fraude, perímetros de medición, marcos temporales y técnicas de estimación. Esto implica que las comparaciones cuantitativas entre fuentes deben interpretarse como indicadores de orden de magnitud, no como mediciones precisas y homogéneas.

En quinto lugar, los tres casos del Capítulo 8 son tipificaciones compuestas construidas a partir de patrones recurrentes documentados en fuentes primarias, no estudios de casos individuales verificables. Si bien esta aproximación es metodológicamente válida para ilustrar dinámicas operativas (Yin, 2018), su valor probatorio es menor que el de casos reales con datos primarios accesibles y replicables.

En sexto lugar, el estudio se centra en plataformas de contenidos digitales en el contexto español y europeo, lo que limita la generalización directa de los resultados y las recomendaciones a otros sectores del ecosistema digital (e-commerce, fintech, servicios de salud digital) y a otras jurisdicciones con marcos regulatorios, estructuras de mercado y niveles de madurez institucional diferentes.

Estas limitaciones, lejos de invalidar las conclusiones del estudio, delimitan su alcance y orientan las líneas de investigación futura desarrolladas en la Sección 11.6.

11.6. Líneas de investigación futura

Las líneas futuras identificadas en este trabajo se orientan a profundizar en el desarrollo de sistemas de detección automatizada multi-capa, el uso defensivo de la inteligencia artificial, la mejora de los marcos de gobernanza y regulación, y el análisis de la dimensión conductual del fraude.

En particular, resulta relevante avanzar en:

- El **desarrollo de modelos predictivos** capaces de anticipar el fraude en tiempo real.
- La **aplicación de inteligencia artificial** explicable en sistemas antifraude.
- El **diseño de mecanismos de cooperación intersectorial**.

- El **estudio de la interacción entre diseño de plataformas y vulnerabilidad del usuario**.

Finalmente, el presente estudio identifica como una dimensión insuficientemente explorada el concepto de fraude moral en el contexto de plataformas de contenido digital. Más allá del daño económico cuantificable, el fraude genera un impacto reputacional, psicológico y social sobre las víctimas que frecuentemente supera la pérdida patrimonial directa: el daño a la dignidad personal, la exposición pública involuntaria, la extorsión mediante contenido íntimo y la manipulación emocional constituyen formas de fraude cuya conceptualización jurídica y técnica requiere un desarrollo específico. Esta línea de trabajo, apenas esbozada en la literatura actual, representa una oportunidad de investigación futura con implicaciones directas para la regulación, la protección de víctimas y el diseño de mecanismos de reparación integral.

Referencias bibliográficas

A. Libros, capítulos de libro y artículos académicos

Akoglu, L., Tong, H., & Koutra, D. (2015). Graph based anomaly detection and description: A survey. Data Mining and Knowledge Discovery, 29(3), 626–688. https://doi.org/10.1007/s10618-014-0365-y

Anderson, R. (2020). Security engineering: A guide to building dependable distributed systems (3.ª ed.). Wiley.

Baxter, G., & Sommerville, I. (2011). Socio-technical systems: From design methods to systems engineering. Interacting with Computers, 23(1), 4–17. https://doi.org/10.1016/j.intcom.2010.07.003

Bostrom, R., & Heinen, J. (1977). MIS problems and failures: A socio-technical perspective. MIS Quarterly, 1(3), 17–32.

Casey, E. (2011). Digital evidence and computer crime: Forensic science, computers, and the internet (3.ª ed.). Academic Press.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM Computing Surveys, 41(3), 1–58. https://doi.org/10.1145/1541880.1541882

Cialdini, R. B. (2009). Influence: Science and practice (5.ª ed.). Pearson.

Cohen, L. E., & Felson, M. (1979). Social change and crime rate trends: A routine activity approach. *American Sociological Review*, 44(4), 588–608. https://doi.org/10.2307/2094589

Cross, C. (2018). Crime, victims and policy: International contexts, local experiences. Palgrave Macmillan.

Cressey, D. R. (1953). *Other people's money: A study in the social psychology of embezzlement*. Free Press.

Gillespie, T. (2018). Custodians of the Internet. Yale University Press.

Gudjonsson, K. (2012). Mastering the super timeline with log2timeline. SANS Digital Forensics. https://www.sans.org

Hadnagy, C. (2018). Social engineering: The science of human hacking (2.ª ed.). Wiley.

Hair, J. F., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. (2017). *A primer on partial least squares structural equation modeling (PLS-SEM)* (2.ª ed.). SAGE.

Levi, M., & Smith, R. (2021). Fraud and its relationship to organized crime. Routledge.

Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. arXiv preprint.https://arxiv.org/abs/1009.6119

Searles, A., Nakatsuka, Y., Ozturk, E., Paverd, A., Tsudik, G., & Enkoji, A. (2023). An empirical study & evaluation of modern CAPTCHAs. En *32nd USENIX Security Symposium (USENIX Security 23)*, 3081–3097. USENIX Association. https://www.usenix.org/conference/usenixsecurity23/presentation/searles

Wardle, C., & Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policymaking. Council of Europe.

Wolfe, D. T., & Hermanson, D. R. (2004). The fraud diamond: Considering the four elements of fraud. *The CPA Journal*, 74(12), 38–42.

Yin, R. K. (2018). Case study research and applications: Design and methods (6.ª ed.). SAGE.

B. Informes institucionales, sectoriales y corporativos

Agencia Española de Protección de Datos. (2021). *Guía para el cumplimiento del deber de informar*. https://www.aepd.es/sites/default/files/2021-10/guia-deber-informar.pdf

Banco Central Europeo [BCE]. (2022). TIBER-EU Knowledge Centre: Lessons learned from the programme (2017–2022). BCE. https://www.ecb.europa.eu/paym/cyber-resilience/tiber-eu/html/index.en.html

Banco de España. (2025). Informe y recomendaciones sobre fraudes en los medios de pago. Banco de España. https://www.bde.es/wbe/es/inicio/noticias/informe-y-recomendaciones-sobre-fraudes-en-los-pagos.html

BioCatch. (2024). Mule account detection and the fight against money laundering. BioCatch Financial Consortium. https://www.biocatch.com/mule-account-detection

BioCatch. (2025). Behavioral biometric signals in financial fraud: Mule account detection update. BioCatch Financial Consortium. https://www.biocatch.com/mule-account-detection

Centro Criptológico Nacional [CCN-CERT]. (2022). Esquema Nacional de Seguridad: Guía de implantación. CCN-CERT. https://www.ccn-cert.cni.es

Coalición de Creadores e Industrias de Contenidos. (2025). Observatorio de piratería y hábitos de consumo de contenidos digitales 2024. https://lacoalicion.es/estudios-del-observatorio/

Comisión Europea. (2024). Supervisory reports on systemic risks under the Digital Services Act: Meta and Google VLOP assessments. Dirección General de Redes de Comunicación, Contenido y Tecnología. https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package

Communications Fraud Control Association [CFCA]. (2023). Global telecommunications fraud loss survey. CFCA. https://www.cfca.org

Commission for Communications Regulation [ComReg]. (2025). SMS Sender ID Registry: Implementation and fraud displacement effects. ComReg.

DARPA. (2023). Media Forensics (MediFor) programme: Final evaluation report. Defense Advanced Research Projects Agency. https://www.darpa.mil/program/media-forensics

Departamento de Seguridad Nacional. (2022). *Estrategia Nacional de Ciberseguridad 2022*. https://www.dsn.gob.es/es/estrategias-nacionales/estrategia-nacional-ciberseguridad

ENISA. (2023a). ENISA Threat Landscape 2023. European Union Agency for Cybersecurity. https://www.enisa.europa.eu/publications/enisa-threat-landscape-2023

ENISA. (2023b). Effectiveness of cybersecurity awareness programmes: Review of evidence. European Union Agency for Cybersecurity. https://www.enisa.europa.eu/publications

European Union Agency for Cybersecurity (ENISA). (2023). *NIS2 Directive: Incident reporting guidelines*. https://www.enisa.europa.eu/publications

ENISA. (2023d). Bias in artificial intelligence: Addressing algorithmic discrimination in cybersecurity applications. European Union Agency for Cybersecurity. https://www.enisa.europa.eu/publications

Europol. (2023a). MISP adoption and threat intelligence sharing in EU law enforcement. EC3, European Cybercrime Centre. https://www.europol.europa.eu

Europol. (2023b). Internet Organised Crime Threat Assessment (IOCTA) 2023. European Union Agency for Law Enforcement Cooperation. https://www.europol.europa.eu/publications-events/main-reports/iocta-report

Europol. (2025). Internet Organised Crime Threat Assessment 2025 (IOCTA). Europol. https://www.europol.europa.eu/cms/sites/default/files/documents/Steal-deal-repeat-IOCTA_2025.pdf

Facebook [Meta]. (2018). Removing coordinated inauthentic behavior from Facebook. Meta Newsroom. https://about.fb.com/news/2018/12/inside-feed-coordinated-inauthentic-behavior/

Global Initiative Against Transnational Organized Crime [GI-TOC]. (2026, 16 de marzo). A world of deceit: Mapping the landscape of the global scam centre phenomenon. GI-TOC. https://globalinitiative.net/wp-content/uploads/2026/03/Kristina-Amerhauser-Alex-Goodwin-A-world-of-deceit-Mapping-the-landscape-of-the-global-scam-centre-phenomenom-GI-TOC-March-2026.pdf

Google. (2023). Advancing account protection: The impact of two-step verification and passkeys. Google Security Blog. https://security.googleblog.com/2023

GSM Association. (2025). *Fraud Manual (Version 22.0)*. [Documento interno GSMA]

GSMA Fraud and Security Architecture Group [FSAG]. (2025). Enhancing controls, empowering the ecosystem. GSMA. [Documento de acceso restringido a miembros del GSMA]

GSMA Fraud and Security Group [FASG#33]. (2025, noviembre). Actas y documentos técnicos del 33.º encuentro del grupo de fraude y seguridad. GSMA, Túnez. [Documentos de acceso restringido a miembros del GSMA]

GSMA Fraud and Security Group [FASG#34]. (2026, febrero). Global Fraud Loss Survey 2025; SMS Blaster Attacks; SIM Swap and Charge to Bill Fraud Patterns. GSMA, Málaga. [Documentos de acceso restringido a miembros del GSMA]

Home Office [Reino Unido]. (2024). Online Safety Act 2023: Mandatory fraud reporting scheme – First year review. Her Majesty's Government. https://www.gov.uk/government/publications

IAON [Gobierno de Aragón, Microsoft, Ibercaja y Fundación Ibercaja]. (2026, 14 de enero). Los deepfakes como un problema real en la sociedad. https://www.ia-on.es/tendencias/los-deepfakes-en-la-sociedad/

ISMS Forum. (2026). Deepfakes: riesgos, casos reales y desafíos en la era de la IA [PDF]. ISMS Forum. https://www.ismsforum.es/ficheros/descargas/deepfake-final1742458135.pdf

Kaspersky. (2023). SpyLoan: Malware detrás de las apps de préstamos con extorsión en América Latina y Europa [comunicado de prensa]. Kaspersky LATAM. https://latam.kaspersky.com/about/press-releases

Kaspersky. (2024, 25 de enero). A todo gas: cómo engañan los estafadores a los «inversores» en internet. Kaspersky Blog. https://www.kaspersky.es/blog/online-investment-dangerous-apps/29557/

Kaspersky. (2025, 5 de septiembre). IT threat evolution in Q2 2025: Mobile statistics. Kaspersky SecureList. https://securelist.com/malware-report-q2-2025-mobile-statistics/117349/

Kaspersky. (2026, 20 de febrero). Phishing y spam: las campañas más descabelladas de 2025. Kaspersky Blog. https://www.kaspersky.es/blog/spam-and-phishing-2025/31858/

Keepnet Labs. (2026). Deepfake statistics and trends 2026: Growth, risks and future insights. Keepnet Labs. https://keepnetlabs.com/blog/deepfake-statistics-and-trends

McAfee. (2024, 25 de noviembre). SpyLoan: una amenaza global que abusa de la ingeniería social. McAfee Labs. https://www.mcafee.com/blogs/es-es/other-blogs/mcafee-labs/spyloan-una-amenaza-global-que-abusa-de-la-ingenieria-social/

McAfee. (2025, 19 de octubre). La alarmante realidad detrás de la escalada del fraude digital en 2025. McAfee. https://www.mcafee.com/blogs/es-es/internet-security/la-alarmante-realidad-detras-de-la-escalada-del-fraude-digital-en-2025/

McAfee. (2026, 8 de enero). El año pasado en estafas: una retrospectiva de 2025 y un adelanto de 2026. McAfee. https://www.mcafee.com/blogs/es-es/security-news/el-ano-pasado-en-estafas-una-retrospectiva-de-2025-y-un-adelanto-de-2026/

Meta. (2023). Coordinated inauthentic behavior report: Q4 2023. Meta Transparency Centre. https://transparency.meta.com/metasecurity/threat-reporting/

Ministerio del Interior [España]. (2026, 5 de febrero). Investigadas doce personas por hacer de mulas bancarias de una organización criminal internacional [nota de prensa]. Ministerio del Interior. https://www.interior.gob.es/opencms/va/detalle/articulo/Investigadas-doce-personas-por-hacer-de-mulas-bancarias-de-una-organizacion-criminal-internacional/

National Cyber Security Centre. (2023a). *Annual review 2023: Active Cyber Defence*. https://www.ncsc.gov.uk/collection/annual-review-2023

National Cyber Security Centre [NCSC]. (2023b). Cyber Aware campaign: 2022–2023 evaluation report. NCSC UK. https://www.ncsc.gov.uk/cyberaware

Netflix Technology Blog. (2022). Improving account security with adaptive multi-factor authentication. Netflix Technology Blog. https://netflixtechblog.com

Omdia. (2025). TV & video industry developments impact brief – June 2025. TechTarget/Omdia.

Singapore Police Force. (2024). Annual scams and cybercrime brief 2023. Singapore Police Force. https://www.police.gov.sg/media-room/statistics

Spotify. (2022). Loud and clear: Trust and safety report 2022. Spotify Technology S.A. https://loudandclear.byspotify.com

UK Finance. (2023). Fraud Managed Service: Privacy-preserving matching technical specification. UK Finance. https://www.ukfinance.org.uk

UNODC. (2013). Comprehensive study on cybercrime. United Nations Office on Drugs and Crime. https://www.unodc.org/documents/organized-crime/UNODC_CCPCJ_EG.4_2013/CYBERCRIME_STUDY_210213.pdf

Visa España. (2025, 22 de diciembre). El 44 % de los españoles declara que le cuesta distinguir contenido auténtico del creado por IA. Visa Inc. https://www.visa.es/sobre-la-corporacion-visa/sala-de-prensa-de-visa/press-releases.3423105.html

YouTube. (2023). YouTube transparency report: Removing fake engagement. Google LLC. https://transparencyreport.google.com/youtube-policy/removals

C. Legislación y normativa

Gobierno de España. (2023). Código Penal (Ley Orgánica 10/1995, de 23 de noviembre, en su versión consolidada 2023). Ministerio de Justicia. https://www.boe.es/buscar/pdf/1995/BOE-A-1995-25444-consolidado.pdf

Unión Europea. (2015). Directiva (UE) 2015/2366 del Parlamento Europeo y del Consejo, de 25 de noviembre de 2015, sobre servicios de pago en el mercado interior [PSD2]. Diario Oficial de la Unión Europea, L 337, 35–127.

Unión Europea. (2016). Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales [RGPD]. Diario Oficial de la Unión Europea, L 119, 1–88.

Unión Europea. (2019). *Reglamento (UE) 2019/881 del Parlamento Europeo y del Consejo, de 17 de abril de 2019, relativo a ENISA y a la certificación de la ciberseguridad de las tecnologías de la información y la comunicación (Reglamento de Ciberseguridad)*. Diario Oficial de la Unión Europea, L 151, 15–69.

Unión Europea. (2022a). Directiva (UE) 2022/2555 del Parlamento Europeo y del Consejo, de 14 de diciembre de 2022, relativa a las medidas destinadas a garantizar un elevado nivel común de ciberseguridad en toda la Unión [NIS2]. Diario Oficial de la Unión Europea, L 333, 80–152.

Unión Europea. (2022b). Reglamento (UE) 2022/2065 del Parlamento Europeo y del Consejo, de 19 de octubre de 2022, relativo a un mercado único de servicios digitales [DSA]. Diario Oficial de la Unión Europea, L 277, 1–102.

Unión Europea. (2024). 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]. Diario Oficial de la Unión Europea, L 2024/1689.

D. Fuentes digitales, medios y publicaciones sectoriales

Acelerápyme (Diputación de Jaén). (2024, 3 de abril). Las 8 mejores plataformas de comercio electrónico en España 2024. https://acelerapyme.dipujaen.es/las-8-mejores-plataformas-de-comercio-electronico-en-espana-2024/

Agencia Comma. (2025, 19 de febrero). Redes sociales en 2025: un ecosistema digital con nuevas reglas. https://agenciacomma.com/marketing-digital/redes-sociales-en-2025-un-ecosistema-digital-con-nuevas-reglas/

AI Safety Institute. (2025). Informe internacional sobre seguridad de la IA 2025 [versión ejecutiva española]. AI Safety Institute. https://internationalaisafetyreport.org/sites/default/files/2025-10/international_ai_safety_report_2025_executive_summary_spanish.pdf

Atresmedia. (2025, 29 de diciembre). El fraude digital aumenta un 40 % en España: 1.200 estafas al día. https://www.atresmedia.com/levanta-la-cabeza/buenas-practicas/ciberseguridad/fraude-digital-aumenta-40-espana-1200-estafas-dia-2025_202512306953a7ee22f0db7daf011949.html

Cybersecurity News. (2025). Las 10 marcas más imitadas en ataques de phishing. https://cybersecuritynews.es/las-10-marcas-mas-imitadas-en-ataques-de-phishing/

Cybersecurity News. (2026, 18 de enero). ¿Cuáles son las marcas más suplantadas por ataques de phishing en el último trimestre de 2025? https://cybersecuritynews.es/cuales-son-las-marcas-mas-suplantadas-por-ataques-de-phishing-en-el-ultimo-trimestre-de-2025/

Digital Innovation News. (2026, 11 de febrero). Fraude digital y transacciones seguras: una prioridad ineludible para las empresas en España. https://digitalinnovationnews.es/fraude-digital-y-transacciones-seguras-una-prioridad-ineludible-para-las-empresas-en-espana/

El Debate. (2024, 3 de julio). Radiografía de la desinformación corporativa: ¿cómo afectan las «fake news» a las empresas? El Debate. https://www.eldebate.com/sociedad/20240703/radiografia-desinformacion-corporativa-como-afectan-fake-news-empresas_210308.html

El País. (2025, 25 de agosto). Cuentas mulas en México: un esquema de lavado de dinero que se expande en el sistema bancario. https://elpais.com/mexico/2025-08-25/cuentas-mulas-en-mexico-un-esquema-de-lavado-de-dinero-que-se-expande-en-el-sistema-bancario.html

EN THEC. (2025). Campañas de phishing, fraude y estafa en redes sociales [documento técnico]. https://enthec.com/wp-content/uploads/2025/02/WP-Campanas-Phishing-Fraude-Estafa.pdf

ESET. (2023, 11 de diciembre). Crecieron las aplicaciones de préstamos maliciosas que engañaban y espiaban a usuarios de Android. ESET LiveLabs. https://www.welivesecurity.com/es/investigaciones/app-prestamos-espian-usuarios-android/

ESET. (2025, 20 de agosto). Qué es el phishing: guía completa 2025. https://www.eset.com/latam/blog/cultura-y-seguridad-digital/que-es-phishing-guia-completa-2025/

Europa Press. (2025, 18 de diciembre). El fraude digital cuesta más de 350 millones de euros al año a la economía española. https://www.europapress.es/economia/noticia-fraude-digital-cuesta-mas-350-millones-euros-ano-economia-espanola-20251218110145.html

FinReg360. (2025, 23 de noviembre). El blanqueo de capitales en la era digital: la amenaza de la IA, las redes sociales y los «influencers». https://finreg360.com/el-blanqueo-de-capitales-en-la-era-digital-la-amenaza-de-la-ia-las-redes-sociales-y-los-influencers/

Forbes Business Council. (2024, 13 de noviembre). Digital marketing trends for 2025 and beyond. Forbes. https://www.forbes.com/councils/forbesbusinesscouncil/2024/11/13/digital-marketing-trends-for-2025-and-beyond/

FractalMedia. (2024, 14 de octubre). Tendencias plataformas OTT y streaming 2025. https://fractalmedia.es/tendencias-plataformas-ott-e-industria-del-streaming-2025/

Future Market Insights. (2025, 10 de septiembre). Digital publishing platforms market (2025–2035). https://www.futuremarketinsights.com/reports/digital-publishing-platforms-market

Hispania Segura. (2025). El auge de la piratería: vuelve a máximos históricos. Una Al Día. https://unaaldia.hispasec.com/2025/11/el-auge-de-la-pirateria.html

IDConline. (2025, 1 de abril). Las cuentas mula y cómo incentivan el lavado en Latinoamérica. IDC. https://idconline.mx/corporativo/2025/04/01/las-cuentas-mula-y-como-incentivan-el-lavado-en-latinoamerica

InfoBae. (2025a, 17 de junio). Suplantaciones de perfiles en redes sociales: así debes actuar para evitar ser presa de estafas. https://www.infobae.com/tecno/2025/06/17/suplantaciones-de-perfiles-en-redes-sociales-asi-debes-actuar-para-evitar-ser-presa-de-ciberdelincuentes/

InfoBae. (2025b, 21 de diciembre). Más de 350 millones de euros: el daño del fraude digital sobre la economía española. https://www.infobae.com/espana/2025/12/21/mas-de-350-millones-de-euros-el-dano-del-fraude-digital-sobre-la-economia-espanola/

InnovaOrgen. (2025, 7 de mayo). SEO 2025: Adaptando estrategias al ecosistema multi-plataforma. https://innovaorigen.io/blog/seo-2025-adaptando-estrategias-al-ecosistema-multi-plataforma/

Keepnet Labs. (2026). Deepfake statistics & trends 2026: Key data & insights [en español]. Keepnet Labs. https://keepnetlabs.com/blg2/deepfake-estadisticas-y-tendencias-2026-principales-datos-e-insights-en-espanol/

Limón Publicidad. (2025, 17 de junio). Canales de marketing digital en 2025: ¿cuáles elegir para destacar tu marca? https://limonpublicidad.com/blog/canales-de-marketing-digital-en-2025-cuales-elegir-para-destacar-tu-marca/

Management Society. (2025, 2 de octubre). ¿Cuántas fake news se consumieron en lo que va del 2025? https://www.managementsociety.net/2025/10/02/cuantas-fake-news-se-consumieron-durante-2025/

MiTek Systems. (2025, 19 de noviembre). La suplantación de identidad en 2025: detectar, prevenir y mitigar. https://www.miteksystems.com/es/blog/suplantacion-identidad-2025-consejos

NordVPN. (2026, 10 de febrero). ¿Es Cupido o un fraude? Las estafas románticas aumentan antes del Día de San Valentín. https://nordvpn.com/es/blog/estafas-romanticas-informe/

PressDigital. (2025, 18 de diciembre). El fraude digital cuesta 350 millones de euros al año a la economía española. https://www.pressdigital.es/articulo/economia/2025-12-18/5708612-fraude-digital-cuesta-350-millones-euros-ano-economia-espanola

Ramsey, C. (2024, 12 de junio). Governments must intervene on anti-fraud funding for real-time payments. *Retail Banker International*. https://www.retailbankerinternational.com/comment/governments-must-intervene-on-anti-fraud-funding-for-real-time-payments/

Red Seguridad. (2025, 20 de enero). ¿Cuál fue la mayor estafa deepfake de la historia, que logró robar 24 millones? https://www.redseguridad.com/actualidad/estafa-multimillonaria-deepfake-mayor-historia_20250120.html

ResearchNester. (2026). Tamaño y cuota de mercado del streaming de música 2026–2035. https://www.researchnester.com/es/reports/music-streaming-market/4383

Revista Mercado. (2025, 20 de mayo). US$ 78 mil millones al año: el costo de las fake news en las finanzas. https://revistamercado.do/market-brief/finanzas/us-78-mil-millones-al-ano-el-costo-de-las-fake-news-en-las-finanzas/

Revista Seguridad. (2025, 29 de enero). Casi 2 millones de cuentas de lavado de dinero reportadas en 2024. https://revistaseguridad.cl/2025/01/29/lavado-de-dinero-reportadas/

SecureList (Kaspersky). (2025, 13 de agosto). Nuevas tendencias en el phishing y las estafas: cómo la IA y las redes sociales están cambiando las reglas del juego. https://securelist.lat/new-phishing-and-scam-trends-in-2025/100247/

Sénal News. (2025, 12 de noviembre). América Latina: la piratería alcanza a más de 40 millones de hogares. https://senalnews.com/es/data/america-latina-la-pirateria-alcanza-a-mas-de-40-millones-de-hogares

Sin Embargo. (2025, 7 de julio). La Policía Cibernética alerta sobre apps y plataformas fraudulentas de inversión. https://www.sinembargo.mx/4673574/la-policia-cibernetica-alerta-sobre-apps-y-plataformas-fraudulentas-de-inversion/

Spain Audiovisual Hub. (2025, 23 de septiembre). La Coalición de Creadores e Industrias de Contenidos: Observatorio de la piratería y hábitos de consumo de contenidos digitales 2024. https://spainaudiovisualhub.digital.gob.es/es/actualidad/la-coalicion-de-creadores-e-industrias-de-contenidos---observat

Stereojoint. (2024, 15 de noviembre). Streaming en 2025: plataformas, tendencias y oportunidades para artistas independientes. https://www.stereojoint.com/post/streaming-en-2025-plataformas-tendencias-y-oportunidades-para-artistas-independientes

UI1 – Universidad Isabel I. (2025, 18 de noviembre). Deepfake: el desafío de la verdad en la era digital. https://www.ui1.es/blog-ui1/deepfake-el-desafio-de-la-verdad-en-la-era-digital

World Economic Forum. (2025, 21 de julio). ¿Cuál es el costo real de la desinformación para las empresas? https://es.weforum.org/stories/2025/07/cual-es-el-verdadero-costo-de-la-desinformacion-para-las-empresas/

Xataka. (2026, 7 de enero). Las mejores plataformas de streaming 2026. https://www.xataka.com/basics/mejores-plataformas-streaming

Glosario de acrónimos y siglas

Constructos del modelo MSFPC

Sigla	Significado
MSFPC	Modelo Socio-Técnico del Fraude en Plataformas de Contenido
ASC	*Attack Surface Complexity* — Complejidad de la superficie de ataque
FMS	*Fraud Method Sophistication* — Sofisticación de los métodos de fraude
FTechS	*Fraud Technical Sophistication* — Sofisticación técnica del fraude
MMC	*Method–Mechanism Coupling* — Acoplamiento método-mecanismo
FIz	*Fraud Industrialization* — Industrialización del fraude
DSM	*Defense System Maturity* — Madurez del sistema de defensa
ACA	*Attacker Adaptive Capacity* — Capacidad adaptativa de los atacantes
FI	*Fraud Impact* — Impacto del fraude
UT	*User Trust* — Confianza del usuario
PS	*Platform Sustainability* — Sostenibilidad de la plataforma

Modelo operativo propuesto

Sigla	Significado
PDIP	Prevención, Detección, Intervención, Persecución (modelo de actuación)
PNIA	Plataforma Nacional de Inteligencia Antifraude (propuesta)

Organismos e instituciones

Sigla	Significado
AEPD	Agencia Española de Protección de Datos
AESIA	Agencia Española de Supervisión de la Inteligencia Artificial

Sigla	Significado
BCE	Banco Central Europeo
CCN-CERT	Centro Criptológico Nacional — *Computer Emergency Response Team*
CFCA	*Communications Fraud Control Association*
CNMC	Comisión Nacional de los Mercados y la Competencia
CNMV	Comisión Nacional del Mercado de Valores
CSIRT	*Computer Security Incident Response Team*
DARPA	*Defense Advanced Research Projects Agency*
ENISA	*European Union Agency for Cybersecurity*
FCSE	Fuerzas y Cuerpos de Seguridad del Estado
GSMA	*Global System for Mobile Communications Association*
INCIBE	Instituto Nacional de Ciberseguridad
ISAC	*Information Sharing and Analysis Center*
ONTSI	Observatorio Nacional de Tecnología y Sociedad
OSI	Oficina de Seguridad del Internauta (INCIBE)
UCO	Unidad Central Operativa (Guardia Civil)
UNODC	*United Nations Office on Drugs and Crime*
WEF	*World Economic Forum*

Normativa y regulación

Sigla	Significado
DSA	*Digital Services Act* — Reglamento (UE) 2022/2065 de servicios digitales
DPIA	*Data Protection Impact Assessment* — Evaluación de impacto en protección de datos

Sigla	Significado
ENS	Esquema Nacional de Seguridad
EU AI Act	Reglamento (UE) 2024/1689 de inteligencia artificial
NIS2	*Network and Information Security Directive 2* — Directiva (UE) 2022/2555
PSD2 / PSD3	*Payment Services Directive* — Directiva de servicios de pago (2.ª / 3.ª)
RGPD	Reglamento General de Protección de Datos — Reglamento (UE) 2016/679
SCA	*Strong Customer Authentication* — Autenticación reforzada de clientes
VLOP	*Very Large Online Platform* (plataformas con >45 millones de usuarios en la UE)

Tecnologías y controles de seguridad

Sigla	Significado
CAPTCHA	*Completely Automated Public Turing test to tell Computers and Humans Apart*
DRM	*Digital Rights Management* — Gestión de derechos digitales
MFA	*Multi-Factor Authentication* — Autenticación multifactor
MISP	*Malware Information Sharing Platform*
OTP	*One-Time Password* — Contraseña de un solo uso
SIEM	*Security Information and Event Management*
STIR/SHAKEN	*Secure Telephone Identity Revisited / Signature-based Handling of Asserted information using toKENs*
STIX	*Structured Threat Information eXpression*
TAXII	*Trusted Automated eXchange of Intelligence Information*
TIBER-EU	*Threat Intelligence-Based Ethical Red Teaming* (marco BCE)

Sigla	Significado
UEBA	*User and Entity Behavior Analytics*
WAF	*Web Application Firewall*

Vectores y tipologías de fraude

Sigla	Significado
APP	*Authorised Push Payment* — Pago autorizado por engaño
IRSF	*International Revenue Share Fraud* — Fraude de compartición de ingresos internacionales
PABX	*Private Automatic Branch Exchange* — Centralita privada

Telecomunicaciones y redes

Sigla	Significado
CDR	*Call Detail Record* — Registro de detalle de llamada
DNS	*Domain Name System*
eSIM	*Embedded SIM* — Tarjeta SIM integrada
HLR	*Home Location Register* — Registro de localización base
IPTV	*Internet Protocol Television*
SIM	*Subscriber Identity Module*
SS7	*Signalling System No. 7* — Sistema de señalización n.º 7
VoIP	*Voice over Internet Protocol*

Modelos de negocio de plataformas

Sigla	Significado
AVOD	*Advertising Video on Demand* — Vídeo bajo demanda con publicidad
BVOD	*Broadcaster Video on Demand* — Vídeo bajo demanda de radiodifusor
FAST	*Free Ad-Supported Television* — Televisión gratuita con publicidad
OTT	*Over-The-Top* — Servicios distribuidos sobre internet
SVOD	*Subscription Video on Demand* — Vídeo bajo demanda por suscripción

Análisis forense e inteligencia

Sigla	Significado
AML	*Anti-Money Laundering* — Prevención del blanqueo de capitales
KYC	*Know Your Customer* — Conocimiento del cliente
OSINT	*Open Source Intelligence* — Inteligencia de fuentes abiertas

Métricas

Sigla	Significado
KPI	*Key Performance Indicator* — Indicador clave de rendimiento
MTTD	*Mean Time to Detect* — Tiempo medio de detección
MTTR	*Mean Time to Respond* — Tiempo medio de respuesta
ROC	*Receiver Operating Characteristic* — Curva de característica operativa del receptor