EL PACCTO 2.0

EU-LAC Partnership on justice and security

USO DE LA INTELIGENCIA ARTIFICIAL POR REDES CRIMINALES DE ALTO RIESGO

2025

Edición: Programa EL PACCTO 2.0

Con la dirección y colaboración de:

Marc Reina Tortosa, Senior Executive Manager, EL PACCTO 2.0
Emilie Breyne, Técnica de proyectos, EL PACCTO 2.0

Autor:

Juan Manuel AGUILAR ANTONIO

DOI: 10.5281/zenodo.16750778

Este documento coordinado por:

Expertise France

Diseño:

Carlos Múgica

Edición no comercial. París, septiembre de 2025

Este documento se ha elaborado con ayuda financiera de la Unión Europea. El contenido de esta publicación es responsabilidad del programa EL PACCTO y sus autores, y en ningún caso debe considerarse un reflejo de los dictámenes de la Unión Europea.

ÍNDICE

ACERCA DEL AUTOR

Profesor – investigador en la Facultad de Estudios Superiores Aragón de la Universidad Nacional Autónoma de México (UNAM). Es miembro del Sistema Nacional de Investigadores (SNI), nivel Candidato (2024–2027). Realizó dos estancias posdoctorales en el Centro de Investigaciones sobre América del Norte (CISAN–UNAM) con proyectos centrados en ciberseguridad, inteligencia artificial y tecnologías emergentes. Becario Fulbright-García Robles para el periodo 2025-2026. Es egresado de los programas "Cyber Policy Development" (2019) y "Combating Transnational Threat Networks in the Americas" (2023) del William J. Perry Center, en la Universidad de la Defensa Nacional, en Washington D.C.

Ha impartido conferencias y docencia en instituciones de seguridad pública y nacional como el CESNAV, el IMEESDN, el CNI, la Policía Cibernética de CDMX y centros de formación en Tamaulipas, Chihuahua, Jalisco y Estado de México. En el plano internacional, es parte de la red de conferencistas del INEES (Guatemala). Y ha sido consultor para la Florida International University (FIU), el Instituto de Estudios Estratégicos de Australia (ASPI), la Iniciativa Global contra el Crimen Organizado Transnacional (GITOC) y el Proyecto Mesoamérica y Cooperasür. Ha sido expositor en foros multilaterales como el IGF de la ONU y la GC3B del Global Forum on Cyber Expertise.

ABREVIACIONES

AIID	AI Incident Database (Base de datos de incidentes de IA)
APIs	Application Programming Interface(s) (Interfaz(es) de Programación de Aplicaciones)
APT	Advanced Persistent Threat (Amenaza Persistente Avanzada)
CJNG	Cártel Jalisco Nueva Generación
CSAM	Child Sexual Abuse Material (Material de Abuso Sexual Infantil)
DDoS	Distributed Denial of Service (Denegación de Servicio Distribuido)
DLS	Data Leak Site (Sitio de filtración de datos)
Europol	Agencia de la Unión Europea para la cooperación policial
GANs	Generative Adversarial Networks (Redes Generativas Antagónicas)
GenIA	Inteligencia Artificial Generativa
GITOC	Global Initiative Against Transnational Organized Crime (Iniciativa Global contra el Crimen Organizado Transnacional)
GNET	Global Network on Extremism and Technology
IA	Inteligencia Artificial
Interpol	Organización Internacional de Policía Criminal
IRGC	Islamic Revolutionary Guard Corps (Cuerpo de la Guardia Revolucionaria Islámica)
ISIS	Islamic State of Iraq and Syria (Estado Islámico)
LLaMA	Large Language Model Meta AI
LLMs	Large Language Model(s) (Modelos de Lenguaje de Gran Escala)
MaaS	Malware as a Service
P2P	Peer-to-Peer (Red entre pares)
PAI	Partnership on AI (Asociación por la IA)
RaaS	Ransomware as a Service
SPOC	Single Points of Contact
SQL	Structured Query Language (Lenguaje de Consulta Estructurado)
UIF	Unidad de Inteligencia Financiera
UNICRI	United Nations Interregional Crime and Justice Research Institute
UNODC	Oficina de las Naciones Unidas contra la Droga y el Delito
VPN	Virtual Private Network (Red Privada Virtual)

INTRODUCCIÓN

La ejecución del crimen con inteligencia artificial ya no es un escenario futurista: es una realidad que está reconfigurando el mapa criminal de América Latina. Lejos de los estereotipos de ciencia ficción, las redes delictivas en la región han comenzado a utilizar modelos generativos, algoritmos de automatización y sistemas de segmentación para estafar, extorsionar, manipular, vigilar o incluso gobernar territorios digitales. Esta transformación no ocurre en el vacío, sino en ecosistemas marcados por la desigualdad tecnológica, la fragmentación institucional y la escasa preparación jurídica frente a delitos que se ejecutan con código, no con armas.

Este estudio ofrece un análisis integral sobre el uso de IA por redes criminales de alto riesgo, así como las capacidades institucionales de los Estados latinoamericanos para enfrentarlas. La investigación combina entrevistas con autoridades de nueve países, estudios de caso y una cartografía estratégica de actores, tecnologías y lagunas normativas. El objetivo no es solo describir un fenómeno emergente, sino contribuir al diseño de respuestas regionales informadas, operativas y con enfoque en derechos.

El documento está organizado en dos bloques temáticos. El Bloque 1 presenta el marco conceptual, las definiciones clave, la metodología utilizada para el mapeo y la tipología de usos criminales de la IA, estableciendo distinciones analíticas entre crimen algorítmico, plataformas autónomas y delitos asistidos por inteligencia artificial. El Bloque 2 desarrolla un marco analítico y estructural del fenómeno, describiendo los modelos operativos del crimen organizado con IA, sus niveles de automatización, los patrones de convergencia entre actores tradicionales y digitales, y ofrece un mapeo regional de redes criminales de alto riesgo que ya integran tecnologías de IA en su modus operandi, identificando sus vínculos con economías ilícitas, capacidades técnicas y lógicas territoriales.

Dentro del análisis por tipo de redes criminales que utilizan la IA para cometer delitos se analizan casos representativos organizados por tipo de actor. En este sentido, se incluyen análisis del ISIS, el Cártel de Sinaloa, el CJNG y KK Park, las redes distribuidas o cibercolectivos como FunkSec, Moustapha Sylla, los Yahoo Boys y el Sindicato del Piso 13, así como las plataformas criminales autónomas bajo el modelo Crime-as-a-Service 5.0, como Xanthorox AI, Storm 2139 y los Dark LLMs. Finalmente, se analizan actores paraestatales y proxies geopolíticos que utilizan IA en operaciones híbridas de desinformación, como Cotton Sandstorm (Irán) y el ecosistema Doppelgänger-Matryoshka (Rusia).

El estudio concluye con propuestas de trabajo a modo de hoja de ruta de políticas públicas dividida en cuatro ejes estratégicos: actualización normativa, fortalecimiento institucional, cooperación regional e interinstitucional, y protección de derechos. Lejos de ofrecer una solución única, el estudio busca ser una plataforma de entendimiento común, capaz de anticipar amenazas, cerrar brechas y construir soberanía digital frente al crimen algorítmico.

BLOQUE 1. IA Y EL CRIMEN ORGANIZADO: MARÇO ANALÍTICO Y ESTRUCTURAL

La convergencia entre IA y crimen organizado no puede comprenderse únicamente como un fenómeno de innovación tecnológica aplicada a actividades ilícitas. En realidad, se trata de una transformación estructural en las formas mismas de organización, operación y adaptación de los actores criminales ante un entorno digital crecientemente automatizado, distribuido e impersonal. Este bloque propone una lectura estratégica de ese fenómeno, a partir de marcos conceptuales que permitan entender a la IA no sólo como herramienta instrumental, sino como factor de estructuración del nuevo orden criminal global.

El análisis parte de una premisa clara: la IA no se distribuye de forma homogénea entre las organizaciones criminales, ni es adoptada con los mismos fines, ni a través de las mismas capacidades. Por el contrario, su integración al ecosistema delictivo responde a motivaciones específicas (económicas, sexuales, políticas o insurgentes), se operacionaliza mediante herramientas algorítmicas diferenciadas (LLMs, deepfakes, scrapers, bots, malware adaptativo), y es determinada por la estructura organizativa de cada red delictiva (jerárquica, distribuida, autónoma o paraestatal). Estas tres dimensiones —motivación, tecnología y organización— constituyen el eje de una matriz estratégica de análisis, que permite observar patrones de apropiación tecnológica por parte de distintos actores criminales y anticipar escenarios de evolución.

El bloque se inicia con un examen del papel de la IA como acelerador del crimen digital, mostrando cómo esta tecnología ha dejado de ser auxiliar o periférica para convertirse en núcleo operativo de campañas de suplantación, extorsión, propaganda o sabotaje. Casos paradigmáticos como WormGPT, FraudGPT o plataformas como Xanthorox AI ilustran esta transición hacia formas de crimen autónomo, sin intervención humana directa, y con capacidades de autoajuste táctico mediante aprendizaje automático. La externalización algorítmica del delito marca así un punto de inflexión: ya no se trata solo de digitalizar el crimen, sino de deshumanizar su ejecución.

A continuación, se presenta una clasificación tipológica de las redes criminales potenciadas por IA, diferenciando entre organizaciones jerárquicas tradicionales que han modernizado su logística y control (como el CJNG o ISIS), redes distribuidas que operan como colectivos sin rostro (FunkSec, Storm-2139, Yahoo Boys), plataformas criminales autónomas que ofrecen crimen como servicio (Crime-as-a-Service), y actores paraestatales o

proxies geopolíticos que instrumentalizan la IA para operaciones híbridas de desinformación, injerencia electoral o ciberguerra. Esta tipología funcional permite mapear los distintos grados de sofisticación tecnológica, descentralización operativa y riesgo estratégico que cada actor representa.

Sobre esta base, se articula una matriz integrada de análisis, que vincula tres variables críticas: motivación criminal, herramientas algorítmicas asociadas y forma organizativa. Dicha matriz permite observar, por ejemplo, cómo los actores con motivaciones económicas privilegian herramientas como LLMs, voice cloning y malware adaptativo para fraudes financieros, mientras que los grupos con motivaciones políticas se inclinan por el uso de bots sociales, algoritmos de targeting ideológico o IA multilingüe para campañas de propaganda. A su vez, esta diferenciación está condicionada por la estructura del actor: no es lo mismo una red informal que accede a herramientas open source, que un cartel que contrata brokers tecnológicos o un Estado que financia desarrollos propios.

Para profundizar esta lectura, se introducen tres modelos conceptuales complementarios que permiten entender cómo se estructura la criminalidad asistida por IA: 1) el modelo de gobernanza extralegal, donde organizaciones como el CJNG o el ISIS utilizan la IA para reforzar su soberanía territorial y disciplinar a sus integrantes. 2) el modelo de red distribuida, propuesto por David Wall, que permite explicar el funcionamiento horizontal, informal y temporal de colectivos como FunkSec. Y 3) el modelo de núcleo-periferia, donde una élite tecnológica desarrolla herramientas que son utilizadas por una periferia de ejecutores, replicando una lógica empresarial descentralizada (como en el caso de Storm2139). Estos enfoques no son excluyentes, sino que ofrecen lentes distintos para comprender la pluralidad morfológica del crimen algorítmico contemporáneo.

Finalmente, se identifican tres tendencias emergentes que marcan la transición hacia una criminalidad algorítmica plena: la automatización total del delito, la disolución de la identidad operativa (crimen sin rostro humano), y la emergencia de regímenes de gobernanza criminal autogestionados en entornos digitales cerrados. Estos escenarios no sólo desafían los marcos jurídicos clásicos, sino que erosionan las capacidades de inteligencia, prevención y atribución de las instituciones estatales.

LA INTELIGENCIA ARTIFICIAL COMO ACELERADOR DEL CRIMEN DIGITAL: DE LA AUTOMATIZACIÓN AL CRIMEN AUTÓNOMO

La incorporación de la IA en el ecosistema criminal ha transformado profundamente la escala, eficiencia y anonimato de las actividades ilícitas. A diferencia de las herramientas digitales convencionales, esta tecnología introduce capacidades de autonomía, adaptabilidad y toma de decisiones que desplazan al actor humano del centro operativo, generando lo que se ha denominado una externalización algorítmica del crimen[1].

Inicialmente, la IA se utilizó como tecnología auxiliar: para automatizar tareas repetitivas como el phishing o el *scraping* de información sensible. Sin embargo, en la actualidad se encuentra en una fase de aceleración estructural, permitiendo ejecutar operaciones delictivas completas con una mínima intervención humana. Casos como *WormGPT* o *FraudGPT*, modelos entrenados explícitamente para generar contenido malicioso, ejemplifican esta evolución hacia un modelo de crimen como servicio algorítmico[2].

El paso del crimen digital tradicional al crimen digital autónomo se manifiesta en distintos niveles. Desde campañas de desinformación generadas en tiempo real por agentes generativos entrenados para polarizar, hasta malware que adapta su comportamiento mediante aprendizaje automático para evadir controles de seguridad[3], la IA no solo automatiza la ejecución delictiva: la optimiza en términos operativos, cognitivos y tácticos[4].

Esta transformación introduce también nuevos dilemas éticos y jurídicos, como la difusa imputabilidad de actos delictivos cometidos por agentes no humanos. En casos donde una IA haya generado y ejecutado acciones ilícitas sin intervención humana directa, las categorías legales tradicionales resultan insuficientes para determinar responsabilidades o escalas de penalidad[5].

CAMBIOS RECIENTES EN LA DISPONIBILIDAD DE HERRAMIENTAS

El auge de los modelos de lenguaje de gran escala (LLMs) como GPT-4, Claude, y LLaMA ha democratizado el acceso a capacidades computacionales avanzadas. Herramientas que antes requerían conocimientos de programación ahora ofrecen interfaces accesibles, lo que permite a actores criminales con escasa formación técnica desarrollar campañas de suplantación de identidad, ingeniería social y estafas financieras con altos niveles de personalización y sofisticación[6].

De esta forma, la IA generativa (GenAI) ha amplificado el impacto del delito digital. Deepfakes de voz, imagen o video, generados con facilidad mediante modelos como *Stable Diffusion*, *Descript* o *ElevenLabs*, permiten llevar a cabo extorsiones, fraudes y chantajes en tiempo real. Según la Internet Watch Foundation, en 2023 se detectaron más de 750,000 imágenes sintéticas de abuso infantil, muchas de las cuales fueron comercializadas en foros de la dark web[7].

Por otro lado, el fenómeno de malware-as-a-service (MaaS) evidencia la consolidación de un mercado digital ilícito en el que se comercializan herramientas automatizadas para llevar a cabo ataques cibernéticos[8]. Plataformas como Xanthoros AI, FunkSec o los Dark LLMs operan bajo modelos de afiliación que incluyen el desarrollo, la distribución y la negociación del rescate, replicando una lógica empresarial descentralizada basada en IA adaptativa[9].

La combinación de LLMs, GenAI y MaaS representa un cambio cualitativo en el crimen organizado: no se trata solo de nuevos medios para viejos fines, sino de nuevas formas de criminalidad emergente que desafían las capacidades tradicionales de prevención, atribución y respuesta. Estas transformaciones requieren marcos analíticos y regulatorios que comprendan a la IA no como una simple herramienta, sino como un actor estructurante del nuevo orden criminal digital[10].

Frente al surgimiento de esta amplia oferta de herramientas de GenIA que permiten innovaciones criminales, TRM Labs[11] estableció una tipología evolutiva en la adopción de IA por parte de organizaciones criminales, distinguiendo entre tres fases: *horizon*, *emerging* y *mature*:

- En la fase horizon, se encuentran delitos aún incipientes como el lavado automatizado de activos, donde el uso de IA tiene un alto potencial disruptivo, aunque baja implementación actual.
- En la fase emerging, se sitúan delitos ampliamente operativos, como el phishing automatizado, los fraudes con deepfakes y la producción de CSAM sintético, los cuales presentan un riesgo creciente y expansión transfronteriza.
- Finalmente, la fase mature proyecta escenarios donde agentes IA autónomos ejecutan delitos complejos sin supervisión humana, como el manejo de wallets, ataques a exchanges y manipulación de mercados financieros.

Esta clasificación permite comprender que el crimen algorítmico no es un fenómeno uniforme, sino una curva de madurez tecnológica con riesgos diferenciados que requieren respuestas reguladoras escalonadas y colaborativas.

1 Caldwell, M., Andrews, J.T.A., Tanay, T., Griffin, L.D. (2020). AI-enabled future crime. Crime Science 9, 14. https://doi.org/10.1186/s40163-020-00123-8.

2 TRM Labs. (2025). The rise of AI-enabled crime: Exploring the evolution, risks, and responses to AI-powered criminal enterprises. https://www.trmlabs.com/resources/blog/the-rise-of-ai-enabled-crime-exploring-the-evolution-risks-and-responses-to-ai-powered-criminal-enterprises

3 Wall, D.S. (2015). Dis-organised crime: Towards a distributed model of the organization of cybercrime. The European Review of Organised Crime 2, 71–90. https://ssrn.com/abstract=2677113.

4 Aguilar Antonio, J.M. (2024). Ransomware gangs and hacktivists: Cyber threats to governments in Latin America. Florida International University, Jack D. Gordon Institute for Public Policy. https://digitalcommons.fiu.edu/jgi_research/65

5 Partnership on AI. (2022). Report on algorithmic risk assessment tools in the U.S. criminal justice system. https://partnershiponai.org/paper/report-on-machine-learning-in-risk-assessment-tools-in-the-u-s-criminal-justice-system/

6 Europol. (2024). Decoding the EU's most threatening criminal networks. Publications Office of the European Union. https://data.europa.eu/doi/10.2813/811566.

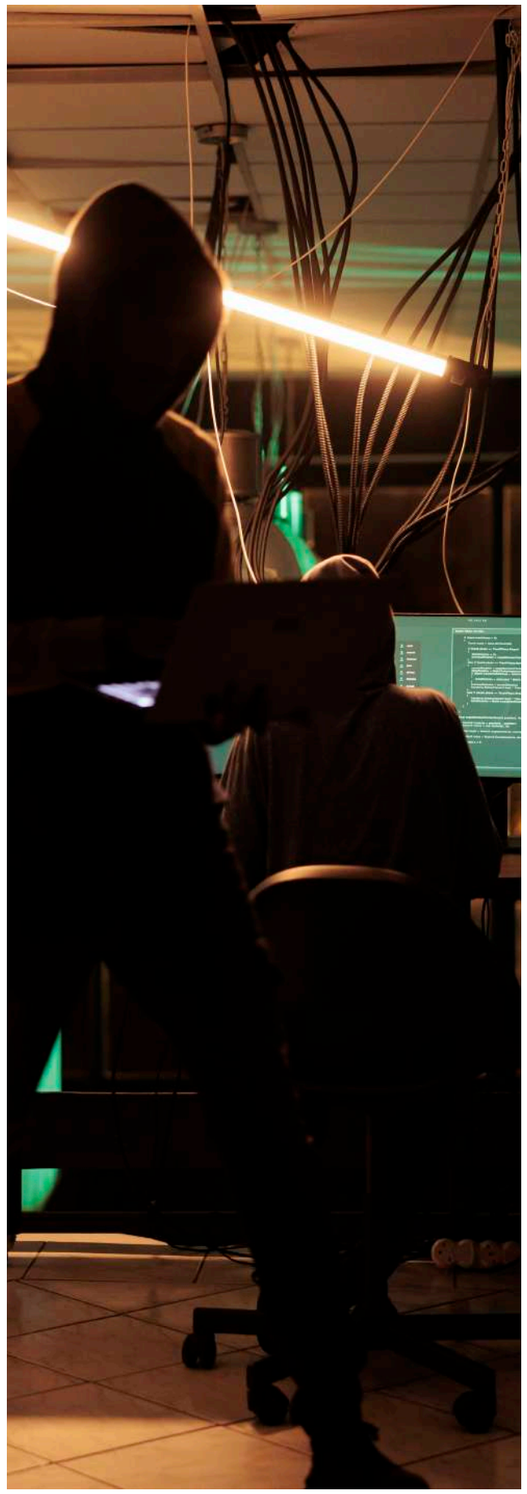
7 TRM Labs. (2025). The rise of AI-enabled crime: Exploring the evolution, risks, and responses to AI-powered criminal enterprises. https://www.trmlabs.com/resources/blog/the-rise-of-ai-enabled-crime-exploring-the-evolution-risks-and-responses-to-ai-powered-criminal-enterprises

8 Aguilar Antonio, J.M. (2024). Ransomware gangs and hacktivists: Cyber threats to governments in Latin America. Florida International University, Jack D. Gordon Institute for Public Policy.

9 Whelan, C., Bright, D., Martin, J. (2024). Reconceptualising organised (cyber)crime: The case of ransomware. Journal of Criminology 57, 45–61. https://doi.org/10.1177/26338076231199793

10 Racoveanu, C. (2024). Artificial intelligence – a double-edged sword: Organized crime's AI vs law enforcement's AI. In Proceedings of the 18th International Conference on Business Excellence, 408–419. ASE Publishing. https://doi.org/10.2478/picbe-2024-0044.

11 TRM Labs. (2025). The rise of AI-enabled crime: Exploring the evolution, risks, and responses to AI-powered criminal enterprises. https://www.trmlabs.com/resources/blog/the-rise-of-ai-enabled-crime-exploring-the-evolution-risks-and-responses-to-ai-powered-criminal-enterprises



TIPOLOGÍAS DE REDES CRIMINALES POTENCIADAS POR IA

CLASIFICACIÓN POR TIPO DE ORGANIZACIÓN CRIMINAL

La integración de IA en las actividades de organizaciones criminales no responde a un único patrón organizativo. Lejos de limitarse a estructuras mafiosas tradicionales, el ecosistema delictivo actual ha incorporado una amplia gama de actores que van desde redes jerárquicas tradicionales en proceso de modernización tecnológica, hasta plataformas criminales autónomas operadas sin intervención humana directa. En esta sección se propone una clasificación basada en la morfología, grado de digitalización y autonomía operativa de los actores criminales que emplean IA para cometer delitos.

- ***Organizaciones jerárquicas tradicionales**:* Estas estructuras, típicamente asociadas con el crimen organizado clásico (carteles de droga, mafias territoriales o redes de tráfico humano), han comenzado a integrar herramientas de IA para optimizar sus operaciones. El Cártel de Sinaloa y el Cártel Jalisco Nueva Generación (CJNG), por ejemplo, han incorporado algoritmos para planificar rutas de tráfico mediante *smart routing*, clonar voces para extorsión afectiva y utilizar herramientas de IA generativa para campañas de phishing financiero[12]. A pesar de conservar una estructura vertical y territorial, estas organizaciones han tercerizado servicios digitales y se apoyan en *brokers tecnológicos* para implementar soluciones algorítmicas específicas[13].

12 Orgaz, C.J. (2024, October 4). Artificial intelligence: 6 ways Latin American criminal groups use AI to commit crimes. BBC News Mundo. https://www.bbc.com/mundo/articles/crej5gwllvlo

13 Europol. (2024). Decoding the EU's most threatening criminal networks. Publications Office of the European Union. https://data.europa.eu/doi/10.2813/811566.

- ***Redes distribuidas y cooperativas de delito basado en tecnología**:* A diferencia de las organizaciones tradicionales, estas redes funcionan sin una jerarquía fija y operan como comunidades colaborativas con nodos semiautónomos. Grupos como FunkSec, Yahoo Boys o el grupo delictivo relacionado con la Operación Cumberland se constituyen más como redes transnacionales que como mafias centralizadas. Sin embargo, esto no evita que sus miembros compartan recursos, técnicas y campañas de ataque. La IA cumple un papel esencial en la generación automatizada de contenido, la extracción de datos mediante scrapers y la diseminación de filtraciones a través de algoritmos de segmentación[14]. Estas redes suelen operar en foros cerrados, redes federadas o servidores autogestionados, pueden desaparecer y reconfigurarse con facilidad, lo que dificulta su atribución legal.

- ***Plataformas criminales autónomas y agentes algorítmicos**:* El surgimiento de sistemas como Xanthorox AI, una plataforma ofensiva capaz de ejecutar ataques cibernéticos sin control humano directo introduce una categoría nueva: la organización criminal sin rostro humano. Estas plataformas operan como *crime-as-a-service*, ofreciendo funciones como redacción de phishing, ejecución de ataques DDoS, penetración de sistemas, análisis de vulnerabilidades, entre otros[15]. La modularidad, escalabilidad y anonimato de estas infraestructuras permiten su uso por parte de múltiples actores delictivos sin intermediación visible. En muchos casos, ni siquiera existe un equipo humano estable detrás, sino un ecosistema algorítmico que interactúa con APIs, wallets y scripts automatizados[16].

- ***Actores paraestatales y proxies geopolíticos**:* Por último, es necesario destacar a los actores vinculados directa o indirectamente con intereses estatales, como Cotton Sandstorm (Irán), Doppelgänger, Storm-1516 o Matryoshka (Rusia). Estos grupos emplean IA para campañas de influencia política, desinformación masiva y ataques cibernéticos con fines estratégicos. Aunque operan bajo una lógica de inteligencia estatal, muchas veces utilizan las mismas herramientas que las organizaciones criminales, como LLMs para manipular narrativas, bots sociales para amplificación de contenido o algoritmos de targeting para segmentación electoral[17]. Su existencia confirma la creciente hibridación entre crimen organizado, operaciones de influencia y ciberguerra.

14 UNODC. (2022). Digest of cyber organized crime: Second edition. United Nations. https://www.unodc.org/unodc/en/cybercrime/global-programme-cybercrime.html

15 Whelan, C., Bright, D., Martin, J. (2024). Reconceptualising organised (cyber)crime: The case of ransomware. Journal of Criminology 57, 45–61. https://doi.org/10.1177/26338076231199793

16 Racoveanu, C. (2024). Artificial intelligence – a double-edged sword: Organized crime's AI vs law enforcement's AI. In Proceedings of the 18th International Conference on Business Excellence, 408–419. ASE Publishing. https://doi.org/10.2478/picbe-2024-0044

17 Europol. (2024). Decoding the EU's most threatening criminal networks. Publications Office of the European Union. https://data.europa.eu/doi/10.2813/811566.

CLASIFICACIÓN POR TIPO DE DELITO FACILITADO POR IA

El despliegue de IA por parte de actores criminales también ha dado lugar a una diversificación significativa de los delitos cometidos mediante medios algorítmicos. Esta sección presenta una clasificación funcional de los principales tipos de delitos potenciados por IA, identificando las organizaciones criminales implicadas y las herramientas utilizadas en cada caso. La evidencia demuestra que el uso de IA no se restringe a un campo delictivo específico, sino que actúa como tecnología transversal capaz de ser aplicada en múltiples fases de la cadena criminal.

- ***Suplantación, deepfakes y fraude algorítmico**:* Organizaciones criminales como Storm-2139, los Yahoo Boys, en Nigeria, y redes regionales en América Latina como el Clan San Roque o los Monstadeudas han explotado herramientas de GenIA para crear contenidos sintéticos con fines de engaño, extorsión y fraude económico. El uso de deepfakes de voz y video permite la suplantación de identidad en tiempo real para engañar a familiares, simular secuestros o desviar fondos de empresas mediante estafas tipo CEO fraud[18].

- ***Producción y circulación de material sexual ilícito sintético**:* Redes como Storm-2139 o la identificada en la Operación Cumberland han producido y distribuido material de abuso sexual infantil sintético generado por IA. Esta práctica evita la necesidad de contacto con víctimas físicas, pero mantiene intactos los patrones de explotación, consumo y comercialización. El impacto es particularmente grave debido a la dificultad legal para tipificar la generación de imágenes "ficticias" como delito, a pesar de su contenido explícito[19].

- ***Ransomware, sabotaje y ciberataques automatizados**:* Grupos como FunkSec, Storm-2139 o la plataforma Xanthorox AI utilizan GenIA para automatizar etapas clave en ataques de ransomware: selección de objetivos, evasión de sistemas de detección y negociación del rescate[20]. La IA permite también adaptar el comportamiento del malware a los entornos operativos detectados, lo que incrementa su letalidad y dificulta su contención[21]. Estas redes operan bajo modelos de *crime-as-a-service* (CaaS), extendiendo sus herramientas a afiliados mediante pagos por su uso.

- ***Mercados criminales, ilícitos y logística criminal automatizada**:* Carteles como el CJNG y el Cártel de Sinaloa han comenzado a integrar IA para optimizar rutas de tráfico de personas y drogas, mediante algoritmos de navegación y predicción de riesgo[22]. El uso de smart routing se ha documentado en informes regionales y permite evitar puntos de control, estimar tiempos de cruce y reducir exposición operativa[23]. Estas aplicaciones suelen desarrollarse en colaboración con brokers digitales regionales que ofrecen soluciones tecnológicas como servicio.

- ***Desinformación, targeting político y guerra cognitiva**:* Grupos como Doppelgänger, Matryoshka y Cotton Sandstorm han empleado IA para influir en elecciones, desestabilizar gobiernos y erosionar la confianza pública. A través de bots sociales, generación de fake news, traducciones automatizadas y targeting ideológico, estos actores manipulan la percepción pública de eventos clave[24]. La convergencia entre crimen organizado y propaganda política digital plantea riesgos inéditos para la seguridad nacional y la estabilidad democrática.

- **Vehículos autónomos o semi-autómos**: grupos criminales como el Cártel de Sinaloa o el CJNS en México, o los Gaitanistas (Clan del Golgo) en Colombia utilizan versiones modificadas de vehículos aéreos autónomos para el tráfico de droga en cantidades menores, la recopilación de inteligencia y el control de rutas de tráfico de ilícitos. Además, grupos colombianos han desarrollado semi-sumergibles completamente autónomos controlados mediante satélite para el tráfico de cocaína a gran escala[25]. La evolución actual constata un interés e inversión por parte de grupos criminales en la innovación tecnológica, no solo de hardware sino también de software.

18 Durán San Juan, I. (2024, October 4). This is how cybercriminals use AI to scam people in Latin America: How you can protect yourself. Infobae. https://www.infobae.com/tecno/2024/10/04/asi-es-como-los-cibercriminales-utilizan-la-ia-para-estafar-personas-en-latinoamerica-como-puede-protegerse/

19 AIID. (2025, febrero 26). Incident 958: Europol Operation Cumberland investigates at least 273 suspects in 19 countries for AI-generated child sexual abuse material. https://incidentdatabase.ai/cite/958/

20 AIID. (2025, abril 7). Incident 1015: Reported darknet launch of Xanthorox AI introduces autonomous cyberattack platform. https://incidentdatabase.ai/cite/1015/

21 Whelan, C., Bright, D., Martin, J. (2024). Reconceptualising organised (cyber)crime: The case of ransomware. Journal of Criminology 57, 45–61. https://doi.org/10.1177/26338076231199793

22 Martínez, R. (2024, August 27). This is how the CJNG uses AI to commit fraud and extortion, according to InSight Crime. Infobae. https://www.infobae.com/mexico/2024/08/27/asi-es-como-el-cnjg-utiliza-ia-para-cometer-fraudes-y-extorsiones-segun-insight-crime/

23 Newton, C. (2024, August 26). How AI is transforming organized crime in Latin America. InSight Crime. https://insightcrime.org/es/noticias/cuatro-formas-inteligencia-artificial-transformando-crimen-organizado-america-latina/

24 UNODC. (2022). Digest of cyber organized crime: Second edition. United Nations. https://www.unodc.org/unodc/en/cybercrime/global-programme-cybercrime.html

25 Triana Sánchez, S. (2025, July 03), La Armada de Colombia intercepta un narcosubmarino teledirigido y con una tecnología que dificulta su rastreo. El País. https://elpais.com/america-colombia/2025-07-02/la-armada-de-colombia-intercepta-un-narcosubmarino-teledirigido-y-con-una-tecnologia-que-dificulta-su-rastreo.html

ASPECTOS CLAVE DE ANÁLISIS PARA REDES CRIMINALES: MATRIZ DE MOTIVACIÓN, TECNOLOGÍA UTILIZADA Y ESTRUCTURA ORGANIZATIVA

Un aspecto esencial para categorizar el uso malicioso de la IA por parte de organizaciones criminales es el articular la motivación de los actores. En ese sentido, se presentan tres dimensiones clave de las organizaciones criminales que utilizan IA para la comisión de delitos: a) motivación criminal, a) herramienta algorítmica y c) estructura organizativa, en una matriz que no solo clasifica, sino que explica con profundidad estratégica los patrones de su adopción por parte de diferentes actores. Estas cuatro motivaciones están interconectadas las unas con las otras dado que en muchos casos los datos son el objeto de ataque por grupos criminales quienes los terminan vendiendo (data as a commodity), utilizando para la comisión de otros delitos, o ambos.

MOTIVACIONES CRIMINALES

Una premisa fundamental de la utilización de IA por parte de organizaciones criminales es el hecho de que esta tecnología no es neutral ni es empleada de manera homogénea. Cada actor delictivo responde a una motivación predominante —económica, sexual, política o insurgente— que condiciona tanto su inversión en tecnología como los riesgos que prioriza. Esta diferenciación permite mapear no solo las formas de daño, sino las lógicas de apropiación tecnológica. De esta forma, en la tabla 1 se presenta una clasificación que describe las motivaciones y la descripción de cada una de estas formas identificadas.

Tabla 1. Motivaciones Criminales.

Motivación	Descripción
Económica	Obtener beneficios mediante fraude, extorsión, robo o lavado de dinero.
Sexual	Explotar contenidos sintéticos de carácter sexual, como CSAM generado por IA.
Política / ideológica	Manipular procesos sociales o psicológicos, electorales o estatales mediante campañas algorítmicas.
Terrorista o insurgente	Emplear IA para propaganda, reclutamiento, sabotaje, manipulación de conducta y desestabilización del orden público.

Fuente: Elaboración propia.

Esta clasificación permite distinguir, por ejemplo, cómo un cartel del narcotráfico enfocado en extorsión automatizada opera con herramientas distintas a un actor insurgente que prioriza propaganda digital y evasión táctica.

HERRAMIENTAS ALGORÍTMICAS ASOCIADAS

Cada motivación se vincula con un conjunto específico de herramientas. En ese sentido, es importante mencionar que la IA no se emplea de forma genérica, sino que se seleccionan herramientas en función del tipo de crimen y su objetivo simbólico o financiero. Esta relación está bien expresada en la siguiente tabla:

Tabla 2. Herramientas de IA según Motivación

Motivación	Herramientas IA predominantes
Económica	LLMs para fraude y phishing, clonación de voz, deepfakes de identidad, malware adaptativo, IA para lavado algorítmico.
Sexual	Generadores de imágenes sintéticas (Stable Diffusion), editores de video con GANs, generación facial y corporal.
Política / ideológica	Bots en redes sociales, LLMs con afinación temática, IA para targeting ideológico, contenido polarizante.
Terrorista / insurgente	IA autónoma para ciberoperaciones, creación de propaganda radicalizada, vigilancia predictiva.

Fuente: Elaboración propia.

Este enfoque operativo permite comprender, por ejemplo, por qué el caso de ISIS con *News Harvest* empleó GenIA multilingüe para propaganda, mientras que el CJNG utiliza modelos conversacionales para fraude emocional y extorsión localizada.

ESTRUCTURAS ORGANIZATIVAS VINCULADAS

La tercera dimensión introduce un elemento diferenciador esencial: la estructura organizativa condiciona el uso de la IA. Esto se representa en la siguiente tabla:

Tabla 3. Estructuras Criminales y Aplicaciones de IA

Estructura organizativa	Aplicaciones de IA predominantes
Jerárquica tradicional	Uso estratégico de IA para vigilancia, logística, extorsión y tráfico, a menudo mediado por brokers tecnológicos externos
Red distribuida informal	IA accesible para automatizar tareas criminales individuales: phishing, sextorsión, campañas pequeñas, scraping y contenido personalizado
Plataforma autónoma	IA como núcleo central (o nodo central) del crimen: desarrollo modular de herramientas automatizadas para uso por múltiples afiliados sin contacto humano directo
Actores paraestatales y proxies geopolíticos	IA empleada a gran escala en campañas de desinformación, targeting electoral, sabotaje de infraestructura o espionaje político

Fuente: Elaboración propia.

Esta matriz muestra que la IA no se distribuye de forma homogénea entre los actores criminales. Mientras que redes informales y colectivos distribuidos aprovechan herramientas de acceso abierto, los actores jerárquicos o paraestatales tienden a financiar desarrollos más sofisticados o contratar servicios personalizados (*crime-as-a-service*). A su vez, la motivación delictiva guía qué combinación tecnológica-estratégica será adoptada.

MODELOS TRADICIONALES AMPLIADOS PARA EL ANÁLISIS

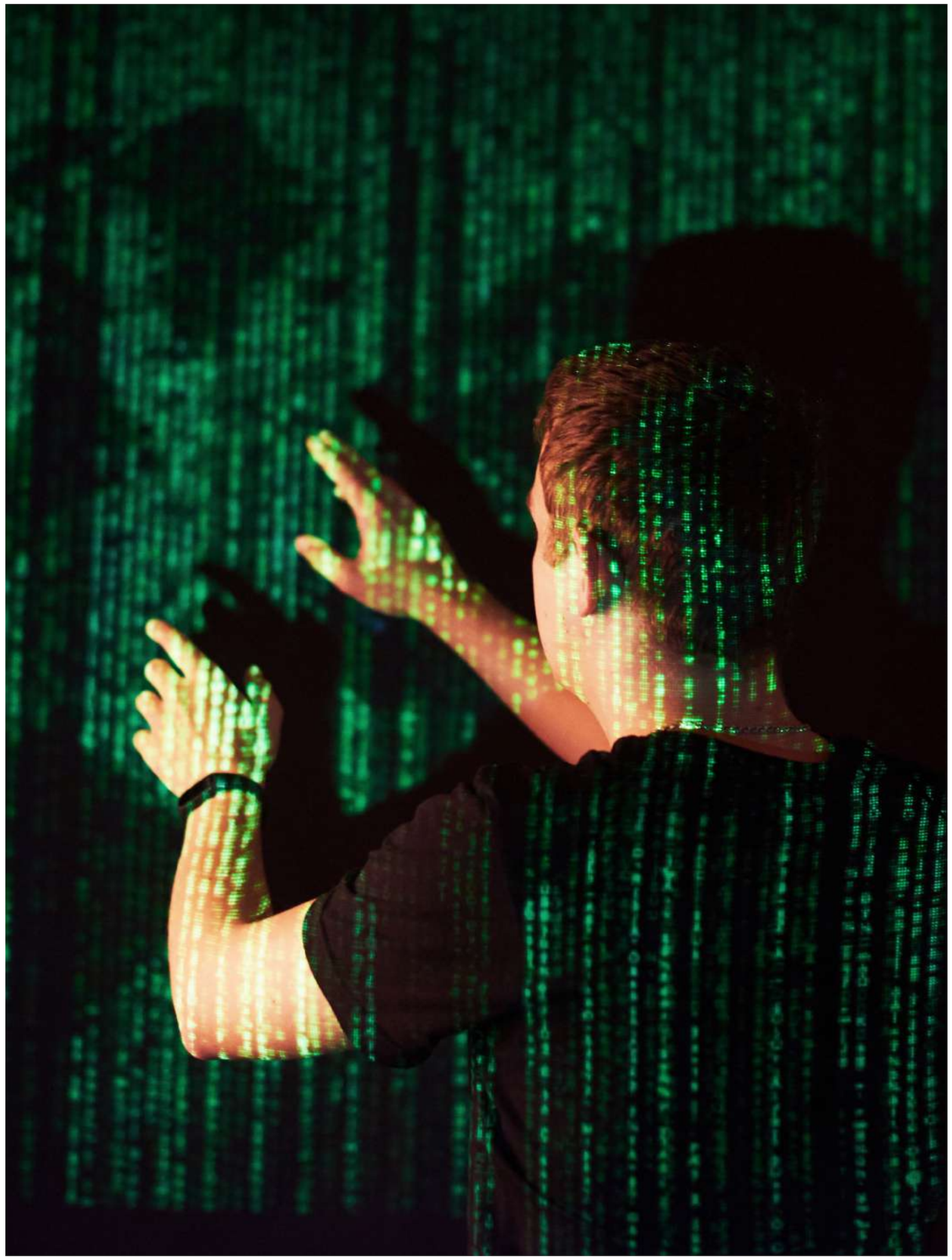
La incorporación de IA en estructuras criminales plantea desafíos significativos para las categorías analíticas tradicionales de las organizaciones delictivas. Comprender a las organizaciones que hoy emplean IA para fines ilícitos requiere revisar y adaptar los marcos teóricos sobre crimen organizado, en función de su grado de estructuración, autonomía tecnológica y forma de gobernanza. En esta sección se sintetizan tres modelos conceptuales complementarios que permiten interpretar el fenómeno: a) el crimen como gobernanza extralegal, b) la noción de crimen desorganizado en redes distribuidas y c) la estructura núcleo-periferia aplicada al delito de alta tecnología.

CRIMEN ORGANIZADO COMO GOBERNANZA EXTRALEGAL (VARESE)

Federico Varese[26] conceptualiza el crimen organizado no solo como una empresa lucrativa, sino como un sistema de gobernanza extralegal que impone reglas, arbitra disputas y regula mercados ilegales o informales donde el Estado está ausente o es ineficaz. Bajo este enfoque, el uso de IA por parte de carteles y mafias no se limita a mejorar la logística o las finanzas criminales, sino que refuerza su capacidad de imponer control territorial, disciplinar actores y ofrecer servicios coercitivos o de protección en mercados paralelos.

Ejemplo de ello es el uso de GenIA por parte del CJNG o el Cartel de Sinaloa para controlar rutas de tráfico de personas o drogas, detectar infiltraciones y aplicar castigos selectivos mediante sistemas de reconocimiento facial. La IA, en este modelo, es un instrumento de soberanía paralela que optimiza la gestión del poder criminal. Este mismo principio se observa en el caso del ISIS, cuya estructura jerárquica ha integrado GenIA para producir propaganda en múltiples idiomas, gestionar la cohesión doctrinal entre células dispersas y automatizar el reclutamiento ideológico, consolidando así un ecosistema de control simbólico transnacional.

26 Varese, F. (2010). What is organised crime? In F. Varese (Ed.), Organized crime: Critical concepts in criminology (Vol. 1, pp. 11–33). Routledge.



De forma aún más radical, el caso de KK Park demuestra cómo mafias chinas y milicias locales han instaurado una forma de gobernanza privada en enclaves como Myawaddy, donde la IA es utilizada para supervisar trabajadores esclavizados, optimizar fraudes globales y aplicar disciplina interna algorítmica, en ausencia total de mecanismos estatales. En estos contextos, la IA no solo facilita el crimen: se convierte en el vector de una nueva arquitectura de dominación informal.

CRIMEN DESORGANIZADO Y REDES DISTRIBUIDAS (WALL)

David Wall[27] propone una crítica al enfoque tradicional que concibe al crimen organizado como jerárquico, vertical y territorial. Su noción de "crimen desorganizado" sugiere que muchas redes criminales en el ciberespacio operan sin liderazgo visible, con estructuras horizontales, temporales y reconfigurables, donde la afinidad funcional sustituye a la lealtad jerárquica. Estas redes se activan y desactivan según oportunidades delictivas, muchas veces mediadas por foros, plataformas o herramientas algorítmicas.

Este modelo permite comprender a colectivos como FunkSec o redes anónimas como Storm-2139, cuya operación se basa en la cooperación descentralizada, la compartición de herramientas como GenIA o scrapers, y la ausencia de jerarquías permanentes. La IA, aquí, es el pegamento técnico que mantiene cohesionada una comunidad informal, donde el crimen se convierte en un proyecto horizontal y distribuido.

ESTRUCTURA NÚCLEO-PERIFERIA Y GOBERNANZA DIGITAL

Whelan, Bright y Martin[28], así como informes de Europol[29] y TRM Labs[30], han propuesto modelos híbridos que describen a los grupos cibercriminales como estructuras núcleo-periferia. En este modelo, un núcleo central tecnológicamente sofisticado desarrolla herramientas, sistemas o servicios (por ejemplo, plataformas de ransomware, bots de targeting, scripts de evasión), mientras que una periferia de afiliados o clientes las utiliza para ejecutar ataques, fraudes o campañas de desinformación.

27 Wall, D.S. (2015). Dis-organised crime: Towards a distributed model of the organization of cybercrime. The European Review of Organised Crime 2, 71–90. https://ssrn.com/abstract=2677113.

28 Whelan, C., Bright, D., Martin, J. (2024). Reconceptualising organised (cyber) crime: The case of ransomware. Journal of Criminology 57, 45–61. https://doi.org/10.1177/26338076231199793.

29 Europol. (2024). Decoding the EU's most threatening criminal networks. Publications Office of the European Union. https://data.europa.eu/doi/10.2813/811566.

30 TRM Labs. (2025). The rise of AI-enabled crime: Exploring the evolution, risks, and responses to AI-powered criminal enterprises. https://www.trmlabs.com/resources/blog/the-rise-of-ai-enabled-crime-exploring-the-evolution-risks-and-responses-to-ai-powered-criminal-enterprises

Este enfoque permite explicar fenómenos como *Ransomware-as-a-Service* (FunkSec) y *Crimeware-as-a-Service* (Xanthorox AI), donde la lógica empresarial digital ha penetrado las formas organizativas del crimen. En lugar de una jerarquía criminal tradicional, observamos ecosistemas de servicios algorítmicos, donde la relación entre núcleo y periferia se rige por contratos, reputación y flujos de capital, no por lealtad o coerción física.

Estos tres modelos no son excluyentes, sino que permiten comprender distintos grados de organización criminal asistida por IA. En algunos casos, como los carteles, predomina la lógica de gobernanza extralegal; en otros, como los colectivos anónimos, prima la lógica distribuida; y en los entornos algorítmicos comerciales, como WormGPT o FraudGPT, predomina la estructura núcleo-periferia.

Adoptar estos marcos conceptuales ofrece una ventaja analítica clave: permite superar el estancamiento de las definiciones normativas tradicionales y comprender la transformación del crimen organizado como un fenómeno en mutación constante, en el que la IA no solo es una herramienta, sino un catalizador de nuevas formas de organización, gobernanza y poder criminal, así como una herramienta de cohesión y mantenimiento de los grupos criminales o individuos que cooperan entre ellos

Tabla 4. Modelos conceptuales para analizar el crimen organizado con IA

Modelo	Estructura organizativa / lógica dominante	Rol de la IA	Ejemplos empíricos
Gobernanza extralegal	*Estructura organizativa:* Jerárquica, con control territorial y coerción física. *Lógica dominante:* Soberanía paralela; imposición de normas	Herramienta para fortalecer dominio territorial y control delictivo	CJNG, ISIS, KK Park
Crimen desorganizado y redes distribuidas	*Estructura organizativa:* Horizontal, informal, temporal *Lógica dominante:* Afinidad funcional; cooperación algorítmica	Medio técnico que mantiene cohesionada una comunidad criminal distribuida	FunkSec, Storm-2139
Núcleo-periferia y gobernanza digital	*Estructura organizativa:* Núcleo central (desarrollo) + periferia de ejecutores *Lógica dominante:* Tercerización algorítmica; crimen como servicio	Infraestructura empresarial criminal, con servicios automatizados por IA	Xanthorox AI, Dark LLMs

Fuente: Elaboración propia.

TENDENCIAS EN LA CONVERGENCIA IA–CRIMINALIDAD ORGANIZADA

La incorporación de la IA en los ecosistemas criminales no constituye simplemente una innovación tecnológica aplicada al delito, sino una mutación estructural en las formas de organización, ejecución y legitimación de las actividades delictivas. A medida que las capacidades algorítmicas se expanden y se hacen accesibles incluso a actores sin conocimientos técnicos avanzados, la IA deja de ser una herramienta auxiliar para convertirse en el núcleo operativo de una nueva criminalidad digital, autónoma y transnacional. Esta convergencia redefine categorías fundamentales como autoría, agencia, territorialidad y trazabilidad, debilitando los marcos jurídicos y estratégicos tradicionales.

La IA permite a actores criminales operar sin rostro, sin cuerpo físico y sin jerarquía visible, ejecutando delitos mediante sistemas autónomos, replicables y adaptativos que circulan por mercados cerrados con lógicas de plataforma. Lo que emerge, entonces, no es una sofisticación más del crimen, sino un nuevo paradigma de criminalidad algorítmica, caracterizado por su opacidad estructural, su capacidad de automatización total y su inserción en arquitecturas de gobernanza paralela.

En ese sentido, se identificaron tres tendencias relevantes que configuran esta transición: a) la automatización completa del delito, b) la disolución de la identidad operativa y c) la aparición de regímenes criminales autorregulados en entornos digitales.

DE LA EXTERNALIZACIÓN DEL DELITO A SU AUTOMATIZACIÓN TOTAL

Las organizaciones criminales han pasado de delegar tareas técnicas (como la producción de malware o el diseño de campañas de phishing) a integrar plataformas autónomas capaces de ejecutar crímenes sin intervención humana directa. Este fenómeno, visible en casos como Xanthorox AI o los Dark LLMs, representa el nacimiento de una economía algorítmica delictiva basada en actores no humanos que ofrecen servicios por suscripción, afiliación o acceso en foros cerrados[31].

31 AIID. (2025, abril 7). Incident 1015: Reported darknet launch of Xanthorox AI introduces autonomous cyberattack platform. AI Incident Database. https://incidentdatabase.ai/cite/1015/

Esta externalización ya no se limita al outsourcing criminal, sino que introduce la figura del crimen automatizado-as-a-service, donde el agente principal es una plataforma de IA operando bajo demanda. Esto impone nuevos desafíos en términos de imputabilidad legal, atribución técnica y respuesta estatal[32].

DESAPARICIÓN DEL ROSTRO HUMANO: CRIMEN SIN IDENTIDAD

La IA ha permitido borrar los rastros visibles de las operaciones criminales. Las redes descentralizadas que emplean IA —como FunkSec, Storm-2139 o bancos de deepfakes sexuales— operan sin rostro, sin voz humana directa y sin jerarquía visible. La clonación de voces, la creación de identidades sintéticas y la falsificación automatizada de documentos han convertido la suplantación algorítmica en una amenaza estructural para la confianza en la identidad, las instituciones y las pruebas digitales[33]. Esta opacidad estructural socava los mecanismos clásicos de inteligencia criminal, ya que el delito no se adscribe a una célula territorial, una familia mafiosa o un liderazgo político, sino a infraestructuras invisibles, transfronterizas y versátiles.

EMERGENCIA DE UNA GOBERNANZA CRIMINAL ALGORÍTMICA

En varios casos documentados, los propios actores delictivos han establecido mecanismos de gobernanza, reputación y arbitraje entre usuarios de herramientas de IA ilícitas. Foros como Exploit.in[34] o RAMP[35] alojan tribunales privados donde se resuelven disputas entre afiliados de plataformas de ransomware, deepfakes-as-a-service o venta de identidades sintéticas. Esto indica una normalización del crimen algorítmico como forma estable de interacción, con códigos, sanciones y jerarquías internas.

Además, ciertos grupos —como Storm-1516 o Cotton Sandstorm— combinan fines ideológicos con tecnologías delictivas, creando híbridos criminales-políticos que operan con lógicas de justicia vigilante, filtración de datos sensibles o intervención simbólica algorítmica. En estos casos, la IA no solo ejecuta delitos, sino que opera como mediadora simbólica del conflicto social, lo que complejiza su clasificación jurídica y estratégica.

En síntesis, la convergencia entre IA y crimen organizado ha generado un nuevo campo de confrontación donde los Estados, las instituciones de justicia y la comunidad internacional enfrentan formas inéditas de amenaza criminal sin precedentes históricos o normativos claros. Esta transformación exige no solo una adaptación técnica, sino una reconstrucción conceptual de lo que entendemos por organización criminal, delito y actor delictivo en el siglo XXI.

32 Racoveanu, C. (2024). Artificial intelligence – a double-edged sword: Organized crime's AI vs law enforcement's AI. In Proceedings of the 18th International Conference on Business Excellence, 408–419. ASE Publishing. https://doi.org/10.2478/picbe-2024-0044.

33 Caldwell, M., Andrews, J.T.A., Tanay, T., Griffin, L.D. (2020). AI-enabled future crime. Crime Science 9, 14. https://doi.org/10.1186/s40163-020-00123-8.

34 Lyngaas, S. (2021, agosto 9). Arbitration among cybercriminals: Inside the underground world of XSS, Exploit and REvil ransomware. CyberScoop. https://cyberscoop.com/arbitration-cybercriminal-xss-exploit-revil-ransomware/

35 SOCRadar. (2023, diciembre 4). *Under the spotlight: RAMP forum*. SOCRadar Threat Intelligence Blog. https://socradar.io/under-the-spotlight-ramp-forum/

METODOLOGÍA DE MAPEO E IDENTIFICACIÓN DE REDES CRIMINALES HERRAMIENTAS

El presente estudio adoptó un enfoque metodológico mixto y adaptativo para la identificación y caracterización de redes criminales de alto riesgo que emplean IA. La acelerada evolución tecnológica, el carácter híbrido de las organizaciones involucradas y la fragmentación de las fuentes disponibles exigieron un modelo de análisis que combinó herramientas de observación sistemática, análisis cualitativo-comparado y una curaduría crítica de literatura gris y bases de datos especializadas. Para ello, se siguieron cuatro pasos fundamentales:

- *Enfoque y fuentes:* Se identificaron y sistematizaron fuentes primarias y secundarias, incluyendo bases de datos de incidentes con IA, literatura académica, informes técnicos y medios de comunicación especializados, priorizando aquellos que ofrecieran trazabilidad y documentación verificable.

- *Criterios de inclusión:* Se establecieron parámetros rigurosos para seleccionar únicamente los casos que evidenciaran el uso comprobado o altamente probable de IA, su vinculación con estructuras delictivas organizadas, y su impacto transnacional o institucional.

- *Limitaciones y consideraciones éticas:* Se analizaron los riesgos de sesgo, sobre interpretación y subregistro derivados de operar con fuentes abiertas. Asimismo, se tomaron precauciones para no amplificar narrativas sensacionalistas.

- *Articulación con el marco conceptual:* Finalmente, cada caso mapeado fue clasificado conforme al marco analítico propuesto en el capítulo anterior, considerando variables como tipo de actor (jerárquico, informal, automatizado), estructura organizativa, tecnologías empleadas y lógica delictiva dominante.

ENFOQUE Y FUENTES

El proceso de mapeo de las redes criminales se fundamentó en un enfoque metodológico mixto, con fuerte énfasis en el análisis documental, comparativo y regional. Dada la complejidad del fenómeno —que combina elementos tecnológicos avanzados, estructuras organizativas opacas y modalidades criminales distribuidas— se recurrió a un conjunto diverso de fuentes primarias y secundarias.

Se privilegió una estrategia de triangulación de datos, combinando bases de datos internacionales, literatura especializada y sistematización manual por región y tipo de actor. Este enfoque permitió no solo verificar la existencia de incidentes criminales facilitados por IA, sino también identificar patrones tecnológicos, actores relevantes y su contexto operativo.

A continuación, se presenta una tabla resumen con los tipos de fuentes utilizadas:

Tabla 5. Tipos de fuentes utilizadas para el mapeo.

Tipo de fuente	Ejemplos representativos	Contribución principal al mapeo
Bases de datos sobre incidentes y vulnerabilidades	AI Incident Database (AIID), Europol, TRM Labs, UNODC	Registro estructurado de casos de uso criminal de IA. Identificación de organizaciones, tecnologías y modus operandi.
Literatura especializada y literatura gris	Informes de Europol, GITOC, TRM; artículos académicos; blogs técnicos (Recorded Future, HackerOne); medios de prensa (BBC, Infobae, The Guardian, etc.)	Análisis contextual, detección de tendencias, triangulación de actores, modalidades delictivas y víctimas.
Sistematización regional de casos y tipologías	Matrices de incidentes, categorización de actores por región (América Latina, UE, Asia), taxonomía tecnológica (LLMs, GenAI, bots, ransomware)	Mapeo de actores relevantes por país; agrupación por tipo de organización (cartel, colectivo, plataforma, Estado).

Fuente: Elaboración propia.

Es importante mencionar, que por sí misma, la fuente más completa para este estudio fue AI Incident Database (AIID), desarrollada por *Partnership on AI*[36]. Esta base recoge incidentes relacionados con el uso de IA que han tenido consecuencias negativas o riesgosas en contextos políticos, económicos, sociales o criminales. Su valor metodológico radica en la sistematización de más de mil entradas organizadas con metadatos técnicos, categorizaciones tipológicas y referencias cruzadas, lo cual facilita la identificación de patrones de uso delictivo de IA, tecnologías empleadas y actores responsables o implicados[37].

Complementariamente, se consultó MITRE's Adversarial Threat Landscape for Artificial-Intelligence Systems (ATLAS), centrada en vulnerabilidades específicas en modelos de IA, útil para comprender vectores de explotación técnica usados por plataformas criminales[38]. Asimismo, reportes técnicos y sistematizaciones de incidentes de cibercrimen con IA fueron extraídos de documentos especializados de Europol[39], TRM Labs[40] y UNODC[41], los cuales permitieron trazar escenarios regionales en sectores críticos como banca, justicia, salud o servicios públicos.

Estas fuentes fueron claves para documentar casos paradigmáticos como Storm-2139 (red global de explotación sexual sintética), Xanthorox AI (plataforma de IA ofensiva modular), Yahoo Boys (red nigeriana de fraudes automatizados), FunkSec (grupo de ransomware como servicio) y Cotton Sandstorm (actor paraestatal vinculado al IRGC[42]).

36 La Partnership on AI (PAI) es una organización sin ánimo de lucro fundada oficialmente el 28 de septiembre de 2016 por grandes empresas de tecnología: Amazon, Facebook (Meta), Google/DeepMind, IBM y Microsoft. . Apple se unió poco después, en enero de 2017.

37 AIID. (2024). AI Incident Database. https://incidentdatabase.ai/

38 MITRE. (2025). ATLAS™: Adversarial Threat Landscape for Artificial-Intelligence Systems. MITRE Corporation. https://atlas.mitre.org/

39 Europol. (2025). EU SOCTA 2025: Strategic report on serious and organised crime in the European Union. Europol.

40 TRM Labs. (2025). The rise of AI-enabled crime. TRM Intelligence Reports.

41 UNODC. (2022). Digest of cyber organized crime: Second edition. United Nations. https://www.unodc.org/unodc/en/cybercrime/global-programme-cybercrime.html

42 Islamic Revolutionary Guard Corps.

LITERATURA ESPECIALIZADA Y LITERATURA GRIS

Se revisaron más de treinta documentos técnicos, informes y artículos académicos producidos entre 2018 y 2025, con énfasis en el vínculo entre crimen organizado e IA. Destacan los informes de situación de Europol y GITOC[43], así como análisis especializados publicados por la Florida International University (FIU)[44]. Estas fuentes proporcionaron un marco interpretativo actualizado sobre la transformación de las organizaciones criminales en entornos digitales y algorítmicos.

Asimismo, se integró literatura gris relevante, proveniente de:

- Blogs especializados: como *Recorded Future, SocRadar* y *HackerOne*, útiles para seguimiento técnico de nuevas herramientas criminales.
- Reportes técnicos de empresas privadas de inteligencia digital: como *TRM Labs*, *Mandiant* y *Check Point Research*.
- Medios periodísticos con enfoque investigativo: tales como *The Guardian*, *Infobae*, *BBC Mundo*, *etc*.

Esta triangulación permitió enriquecer el análisis con diferentes escalas de observación (técnica, institucional y territorial), fortaleciendo la identificación de casos relevantes y tendencias de uso de IA por parte de organizaciones criminales.

43 GITOC. (2023). Global organized crime index 2023. Global Initiative Against Transnational Organized Crime.

44 Aguilar Antonio, J.M. (2024). Ransomware gangs and hacktivists: Cyber threats to governments in Latin America. Research Publications 65. https://digitalcommons.fiu.edu/jgi_research/65

MAPEO TIPOLÓGICO DE ACTORES CRIMINALES: CLASIFICACIÓN FUNCIONAL Y OPERATIVA

Como resultado del proceso de sistematización y clasificación de los casos identificados, se creó una tipología funcional y comparativa de los principales actores criminales que incorporan IA en sus operaciones. A diferencia de una mera enumeración por país o sector, el enfoque adoptado permite observar cómo diferentes morfologías organizativas adaptan las capacidades de la IA a sus objetivos delictivos específicos. La clasificación se organiza en cuatro grandes categorías presentes en la tabla 6. Este marco permite analizar patrones operativos, tecnologías empleadas, beneficiarios y víctimas, proporcionando una base analítica para el diseño de estrategias diferenciadas de respuesta y gobernanza.

Tabla 6. Mapeo de organizaciones y casos relevantes (tipología criminal).

Clasificación	Organización / Casos	Descripción breve del grupo
Organizaciones jerárquicas tradicionales	1. ISIS (News Harvest) 2. Cartel de Sinaloa y Cartel 3. Jalisco Nueva Generación (CJNG) 4. KK Park (Unión Nacional Karen – KNU, redes afiliadas a Wan Kuok-koi, Guardia Fronteriza de Myanmar)	Estructuras verticales, con mando centralizado y control territorial o temático. Usan IA para extender sus capacidades logísticas, financieras o coercitivas.
Redes distribuidas o cibercolectivos	1. FunkSec 2. Yahoo Boys (Nigeria) 3. Montadeudas CDMX 4. Sindicato del Piso 13 de Poipet 5. Operación Cumberland 6. Clan San Roque	Grupos descentralizados, sin jerarquía fija, operan en red abierta. Usan IA para sabotaje digital, ransomware y ataques a infraestructura pública y privada.
Plataformas criminales autónomas (Crime-as-a-Service)	1. Dark LLMs (WormGPT, FraudGPT, DarkBARD) 2. Storm-2139 3. Xanthorox AI	Entornos automatizados que ofrecen servicios criminales digitales (deepfakes, ransomware, malware, bots). Operan con lógica empresarial descentralizada y modular.
Actores paraestatales y proxies geopolíticos	1. Cotton Sandstorm 2. Doppelgänger, Storm-1516, Matryoshka	Actores vinculados a estructuras gubernamentales o militares. Utilizan IA para propaganda, manipulación de elecciones, operaciones de influencia y ciberguerra.

Fuente: Elaboración propia.

Este ejercicio de mapeo demuestra que el uso de IA no responde únicamente a capacidades tecnológicas, sino a una lógica organizativa específica. Cada tipo de actor —desde redes jerárquicas hasta plataformas autónomas— emplea la IA de forma distinta, maximizando sus ventajas comparativas y operativas. Esta clasificación permite establecer líneas analíticas más precisas para el monitoreo, la regulación y la cooperación internacional, atendiendo tanto a la morfología criminal como a los vectores tecnológicos que definen esta nueva arquitectura delictiva global.

BLOQUE 2. MAPEO DE REDES CRIMINALES DE ALTO RIESGO QUE USAN IA

Uno de los mayores desafíos para el análisis contemporáneo del crimen organizado asistido por IA es su opacidad estructural y fluidez operativa. A diferencia de las organizaciones criminales tradicionales, cuyas trayectorias, liderazgos y territorios podían ser rastreados mediante inteligencia convencional, las redes algorítmicas actuales se caracterizan por su versatilidad morfológica, su diseminación transnacional y su capacidad para operar en entornos digitales cerrados, cifrados o efímeros. Ante esta realidad, el presente bloque desarrolla un ejercicio de mapeo estratégico y tipológico, orientado a identificar, clasificar y analizar a los actores criminales de alto riesgo que incorporan tecnologías de IA en sus operaciones ilícitas.

El propósito de este bloque no es simplemente compilar una lista de casos, sino construir una cartografía funcional que permita comprender cómo diferentes formas organizativas —desde carteles tradicionales hasta colectivos distribuidos, plataformas autónomas o proxies estatales— están integrando la IA como parte de sus lógicas delictivas, estructuras de poder y mecanismos de expansión. El enfoque adoptado privilegia la dimensión comparativa, reconociendo que el uso de tecnologías algorítmicas no se distribuye de manera homogénea, sino que responde a contextos regionales, niveles de sofisticación técnica y objetivos criminales divergentes.

Para ello, se desarrolla un mapeo tipológico de los actores criminales, estructurado en cuatro grandes categorías ya descritas en el marco analítico: 1) organizaciones jerárquicas tradicionales, 2) redes distribuidas o cibercolectivos, 3) plataformas criminales autónomas y 4) actores paraestatales y proxies geopolíticos. Cada categoría agrupa a organizaciones o casos relevantes —como el CJNG, FunkSec, Xanthorox AI o Cotton Sandstorm— que ilustran con claridad distintas formas de apropiación algorítmica del crimen.

Este mapeo no solo identifica a los actores, sino que analiza sus tecnologías empleadas, estructuras internas, vectores de ataque y públicos objetivo, lo cual permite detectar patrones operativos y anticipar trayectorias de riesgo. Por ejemplo, mientras que grupos como los Yahoo Boys han utilizado herramientas de GenIA para realizar estafas personalizadas a escala global, el caso de la Operación Cumberland muestra como redes han perfeccionado modelos de producción sintética de material de abuso sexual infantil, sin contacto físico con las víctimas. A su vez, el Cártel de Sinaloa y el CJNG han incorporado algoritmos de navegación predictiva para la logística de tráfico humano y de drogas, y grupos como Doppelgänger han desplegado IA para manipulación informativa con fines estratégicos.

La lógica comparativa que subyace a este mapeo no responde únicamente a un afán clasificatorio, sino a una necesidad operativa: entender cómo los diferentes tipos de actores criminales adoptan tecnologías de IA para maximizar sus capacidades, diversificar sus ingresos, eludir la atribución y perpetuar estructuras de control. Esta comprensión es clave para el diseño de estrategias de contención que no se limiten a reaccionar ante delitos consumados, sino que actúen de manera preventiva, anticipando las formas en que las capacidades algorítmicas transforman el crimen organizado en el siglo XXI.

ORGANIZACIONES JERÁRQUICAS TRADICIONALES

A lo largo de las últimas dos décadas, las organizaciones criminales con estructuras jerárquicas tradicionales han sido erróneamente interpretadas como obsoletas frente a la irrupción de nuevas formas de delito distribuido, cibercolectivos o plataformas de crimen autónomo. Sin embargo, lo que ha ocurrido en realidad es un fenómeno de reconfiguración profunda, donde las viejas formas de autoridad vertical han encontrado en la IA no un sustituto, sino una extensión estratégica de su poder.

Lejos de diluirse, los esquemas de mando centralizado —como los que estructuran a ISIS, los carteles mexicanos o el complejo de KK Park en Myanmar— han incorporado la tecnología como instrumento de control doctrinal, disciplinamiento interno y expansión operativa transnacional. La IA no reemplaza el liderazgo; lo codifica, lo traduce a escala algorítmica y lo proyecta en dominios donde la presencia física ya no es necesaria.

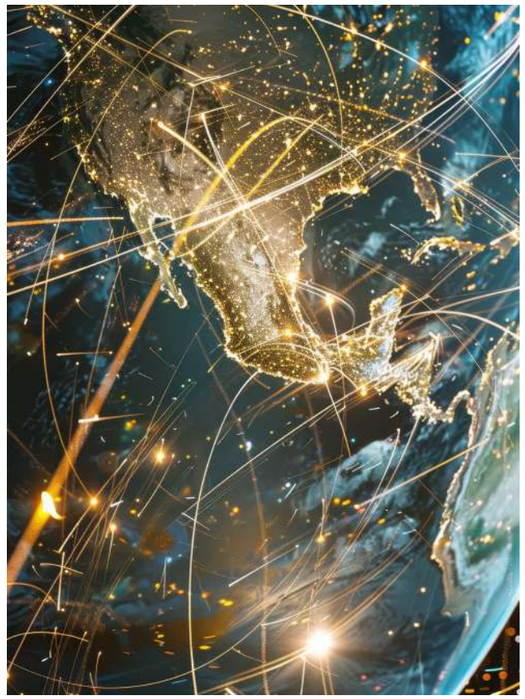
En este nuevo escenario, la centralización del mando emerge como una ventaja adaptativa. Las organizaciones jerárquicas, con cadenas de mando claramente definidas y procesos internos de obediencia estructurada, son capaces de adoptar herramientas tecnológicas de forma más ordenada, disciplinada y eficiente que sus contrapartes descentralizadas. La verticalidad permite que las órdenes se automaticen, que los dispositivos de control sean internalizados por los cuadros medios, y que las nuevas tecnologías —desde generadores de deepfakes hasta dashboards de vigilancia y monitoreo— funcionen como extensiones de un mando legitimado internamente. La eficacia de estas estructuras no está en su flexibilidad, sino en su capacidad de traducir el poder simbólico del jefe en órdenes replicables, predecibles y escalables, sin que eso signifique menor adaptabilidad.

La IA, en este contexto, opera como una tecnología de refuerzo más que de disrupción. En el caso del Estado Islámico, por ejemplo, permite la automatización doctrinal: bots que replican discursos salafistas, cuentas que difunden material pro-yihadista a través de traducciones automáticas, videos generados que mantienen vigente la retórica del martirio. En el caso del CJNG y el Cártel de Sinaloa, la IA se articula con estrategias de propaganda, extorsión inteligente, reconocimiento facial y vigilancia territorial. Y en KK Park, la tecnología se vuelve el esqueleto de una industria criminal algorítmica: un sistema donde el fraude emocional, la estafa financiera y la esclavitud digital se entrelazan gracias a la programación estratégica de explotación transfronteriza. La IA, en todos estos casos, no erosiona el poder; lo hace reproducible.

A pesar de compartir una lógica jerárquica, los modelos que representan ISIS, CJNG/Sinaloa y KK Park no son homogéneos. El primero responde a una estructura teocrática-doctrinal, donde el mando es espiritual pero la organización es profundamente pragmática en su adaptación tecnológica. El segunda encarna una verticalidad militar-comercial híbrida, con jefes de plaza, operadores financieros, comandos armados y un uso instrumental de la violencia algorítmica. El tercero —quizás el más sofisticado— mezcla gobierno empresarial con logística militar y control simbólico digital. Este modelo, basado en la producción industrial del engaño, la coerción emocional y la captura simbólica de las víctimas, representa una mutación del crimen tradicional hacia formas híbridas que merecen ser conceptualizadas como enclaves de gobernanza criminal algorítmica.

Lo que une a estos tres modelos es su capacidad de ejercer poder sin presencia física. En cada uno, la violencia ya no se manifiesta necesariamente como un acto corporal, sino como una arquitectura de daño automatizado. El AK-47 ha sido sustituido, en muchos casos, por scripts, deepfakes o dashboards. La violencia, hoy, se ha digitalizado sin perder efectividad, y se ha vuelto incluso más insidiosa: la IA no mata con balas; lo hace con secuencias de código.

Este desplazamiento del daño físico al simbólico tiene implicaciones institucionales profundas. Si bien el número de muertes puede disminuir en ciertas regiones, el impacto acumulativo sobre la cohesión social, la gobernanza estatal y la percepción de seguridad es devastador. Estas organizaciones jerárquicas no buscan necesariamente el control territorial clásico, sino el control de flujos: flujos de datos, capitales, dinero, etc. En este sentido, el poder que ejercen se acerca



más al de una infraestructura que al de un ejército. KK Park no necesita conquistar ciudades; le basta con operar interfaces que manipulen a víctimas en Brasil, Japón o Canadá desde un edificio invisible en Myanmar. CJNG no necesita estar presente en Nueva York para extorsionar a comerciantes locales. ISIS ya no necesita enviar emisarios; le basta con que su retórica llegue a jóvenes radicalizados a través de Telegram.

Así, las organizaciones criminales jerárquicas están dando un salto cualitativo en su relación con la tecnología. Ya no solo la utilizan como medio; la han convertido en una dimensión estructural de su operación. Sus líderes entienden que el futuro del crimen no está únicamente en la fuerza, sino en la codificación. La verticalidad ya no requiere presencia: puede residir en la programación.

En este bloque, se presentan estos tres casos paradigmáticos donde esta convergencia entre jerarquía y tecnología se manifiesta con mayor claridad. No se trata de organizaciones nuevas, sino de mutaciones avanzadas de poderes ya conocidos, ahora potenciados por la arquitectura algorítmica. El estudio de estos modelos es, por tanto, una advertencia: en el mundo contemporáneo, el crimen organizado ya no necesita esconderse. Puede simplemente digitalizarse, escalarse y operar fortalecido por la IA.

CASO 1. CJNG Y CARTEL DE SINALOA

El uso de IA por parte del CJNG y el Cártel de Sinaloa refleja una evolución organizativa clave en el crimen organizado mexicano. Aunque sus arquitecturas internas difieren —el CJNG funciona con un mando vertical de tipo militar, mientras que el Cártel de Sinaloa se articula a través de redes flexibles y descentralizadas—, ambas organizaciones han convergido en la integración funcional de tecnologías emergentes para trasladar parte de sus operaciones al espacio digital[45].

Esta transición no supone el abandono de la violencia tradicional, sino su reconfiguración. En la actualidad, el control ya no depende exclusivamente del enfrentamiento armado, sino de la capacidad para generar amenazas verosímiles a través de medios digitales, simulando secuestros, suplantando identidades o manipulando emocionalmente a las víctimas mediante tecnologías algorítmicas. Al mismo tiempo, estas organizaciones han comenzado a integrar la lA como infraestructura operativa en otras dimensiones críticas de su funcionamiento: la optimización de cadenas logísticas delictivas, el perfeccionamiento de esquemas de lavado financiero multijurisdiccional y la automatización de procesos operativos internos, lo que les permite reducir tiempos, costos y exposición humana en sus actividades de tráfico, extorsión y blanqueo de activos.

Por su parte, el CJNG, ha centralizado el desarrollo de esquemas de extorsión automatizada con IA generativa. Esta organización ha sido pionera en el uso de clonación de voz y bots conversacionales para realizar fraudes emocionales, como el conocido *pig butchering*, en el que se construyen vínculos afectivos falsos con víctimas a lo largo del tiempo, hasta inducirlas a transferir grandes cantidades de dinero[46].

En paralelo, el Cártel de Sinaloa ha adoptado una lógica de replicación descentralizada. Diversas células operan de manera autónoma en la ejecución de campañas de smishing (suplantación de funcionarios y manipulación de identidades digitales). Este modelo facilita una rápida adopción de herramientas como *deepfakes*, traducción algorítmica o geolocalización automatizada, muchas veces sin necesidad de una estructura de mando centralizada[47]. Su fortaleza radica en la diseminación fragmentada del poder tecnológico, lo que dificulta los esfuerzos de rastreo y neutralización por parte de las autoridades.

Ambos cárteles han entendido que la IA no es solo una herramienta técnica, sino una infraestructura operativa que permite mantener el control simbólico sobre personas, flujos financieros y territorios virtuales. El CJNG lo hace desde una lógica de comando y control; el Cártel de Sinaloa, desde una lógica de adaptabilidad y expansión celular. En síntesis, ambos grupos han reestructurado parte de su aparato operativo en torno a la IA, creando ecosistemas criminales híbridos en los que conviven la violencia tradicional y la coerción algorítmica. Esta transformación exige un nuevo enfoque desde las políticas públicas y la seguridad nacional: ya no basta con neutralizar operativos armados, sino con entender y desactivar las arquitecturas técnicas que sostienen la violencia automatizada.

Tecnologías utilizadas

La digitalización progresiva del crimen organizado en México, específicamente en las estructuras operativas del CJNG y el Cártel de Sinaloa, ha dado lugar a una arquitectura tecnológica delictiva profundamente funcional. Lejos de tratarse de una sofisticación marginal, el uso de IA ha sido absorbido como infraestructura táctica y simbólica, multiplicando la eficiencia de sus operaciones sin requerir exposición física directa.

Entre los componentes centrales de este ecosistema destaca el uso de inteligencia artificial generativa, tanto textual como audiovisual. Estas herramientas permiten la producción automatizada de mensajes persuasivos, redactados en distintos registros emocionales y adaptados culturalmente. La capacidad de simular interacciones conversacionales verosímiles —a través de modelos de lenguaje entrenados para sostener diálogos prolongados— ha sido esencial para el desarrollo de esquemas de fraude afectivo y emocional, como el conocido pig butchering[48].

Un elemento complementario y particularmente perturbador ha sido la incorporación de sistemas de clonación de voz. Mediante estas tecnologías, las organizaciones criminales han logrado replicar voces de familiares o figuras de autoridad con alto nivel de fidelidad acústica, lo que ha permitido montar escenarios de secuestro, emergencia médica o coerción judicial que resultan emocionalmente devastadores para las víctimas[49].

En el plano visual, la utilización de deepfakes ha comenzado a consolidarse como recurso de alto impacto simbólico. Se han identificado casos en los que se emplean tecnologías de reconstrucción facial y lip-sync automatizado para simular videos de supuestas agresiones, capturas o amenazas. Aunque su implementación aún es limitada en volumen, su efectividad psicológica ha sido ampliamente documentada por analistas de seguridad digital[50].

La automatización de la amenaza es otra dimensión clave. Bots conversacionales programados con IA permiten escalar la extorsión sin requerir interlocución humana directa. Estos bots están diseñados para detectar patrones emocionales en las respuestas de la víctima y ajustar su tono o contenido en tiempo real, maximizando la presión psicológica mediante una lógica de retroalimentación algorítmica.

En el terreno financiero, el uso de criptomonedas y tecnología blockchain ha sido crucial para el ocultamiento y desplazamiento transnacional de fondos ilícitos. El Cártel de Sinaloa, en particular, ha estructurado redes de triangulación con casas de cambio clandestinas chinas, permitiendo la conversión de ganancias del tráfico de fentanilo en activos digitales, y su posterior transformación en yuanes a través de corredores paralelos[51]. Estas operaciones no solo evaden el sistema bancario internacional, sino que lo reemplazan con una arquitectura financiera descentralizada, resistente a la trazabilidad institucional.

A este conjunto se suma el uso generalizado de software de anonimato y evasión digital. VPNs comerciales, redes Tor, direcciones IP rotativas, servidores espejo y cuentas desechables permiten la circulación de amenazas, manuales operativos y contenido fraudulento sin dejar huella rastreable. Esta capa de protección técnica no solo garantiza

45 García, S. (2025, May 8). How criminal groups have adapted to the digital age. InSight Crime. https://insightcrime.org/es/noticias/como-grupos-criminales-adaptado-era-digital/

46 Martínez, R. (2024, August 27). This is how the CJNG uses AI to commit fraud and extortion, according to InSight Crime. Infobae. https://www.infobae.com/mexico/2024/08/27/asi-es-como-el-cnjg-utiliza-ia-para-cometer-fraudes-y-extorsiones-segun-insight-crime/

47 Martínez, R. (2024, May 8). These are the apps used by the Sinaloa Cartel and Los Chapitos to communicate without leaving a trace. Infobae. https://www.infobae.com/mexico/2024/05/08/estas-son-las-aplicaciones-que-usan-el-cartel-de-sinaloa-y-los-chapitos-para-comunicarse-sin-dejar-rastro/

48 Martínez, R. (2024, August 27). This is how the CJNG uses AI to commit fraud and extortion, according to InSight Crime. Infobae. https://www.infobae.com/mexico/2024/08/27/asi-es-como-el-cnjg-utiliza-ia-para-cometer-fraudes-y-extorsiones-segun-insight-crime/

49 AIID. (2024). Incident 725: Cartels reportedly using AI to expand operations into financial fraud and human trafficking. https://incidentdatabase.ai/cite/725

50 Newton, C. (2024, August 26). How AI is transforming organized crime in Latin America. InSight Crime. https://insightcrime.org/es/noticias/cuatro-formas-inteligencia-artificial-transformando-crimen-organizado-america-latina/

51 TRM Labs. (2024, July 26). Authorities unravel the Sinaloa Cartel's connection to Chinese money launderers. TRM Blog. https://www.trmlabs.com/es/resources/blog/authorities-unravel-the-sinaloa-cartels-connection-to-chinese-money-launderers

la persistencia del mensaje, sino que incrementa la resiliencia operativa frente a intervenciones estatales.

Finalmente, una de las prácticas más extendidas es el uso de software de scraping y minería de datos, aplicado a redes sociales, directorios públicos y bases de datos filtradas. Esta información es procesada por sistemas clasificatorios que construyen perfiles de vulnerabilidad individual, seleccionando víctimas por edad, localización, nivel educativo o tipo de empleo[52]. El cruce de estas bases con motores de IA ha elevado la precisión de campañas criminales a niveles inéditos.

Modus Operandi

La incorporación de IA por parte del CJNG y el Cártel de Sinaloa no ha derivado en un cambio superficial en sus métodos de acción, sino en una reconfiguración operativa integral que afecta simultáneamente las esferas de coerción, fraude, circulación financiera y control simbólico. A diferencia de insurgencias como ISIS, donde la integración tecnológica sigue una lógica ideológica centralizada, en el caso mexicano la lógica dominante es instrumental y modular, orientada a maximizar la eficacia delictiva con menor exposición organizativa posible.

En el caso del CJNG, el empleo de tecnologías digitales responde a un diseño centralizado. Esta organización ha estructurado unidades tecnológicas especializadas, responsables de ejecutar campañas de fraude emocional y extorsión automatizada desde servidores seguros y bajo protección de redes cifradas. Esta relación se construye progresivamente, alimentada por inputs emocionales y psicológicos de la víctima, hasta inducir una transferencia económica voluntaria que puede prolongarse durante semanas o meses[53].

A este esquema se suma una segunda modalidad: la extorsión por suplantación de identidad, en la que se utilizan voces clonadas de familiares cercanos —obtenidas a partir de mensajes de voz, redes sociales o grabaciones previas— para simular secuestros o situaciones críticas. Estas llamadas, reproducidas mediante plataformas de síntesis vocal, son acompañadas frecuentemente por mensajes multimedia con deepfakes, que simulan escenas de violencia o cautiverio. Esta combinación maximiza el efecto emocional y reduce la capacidad de discernimiento de la víctima.

Ambos cárteles han integrado el uso de mensajería automatizada y distribución algorítmica de amenazas. Las campañas de smishing y vishing, enviadas desde cuentas temporales o mediante redes cifradas, son alimentadas con datos personales obtenidos por scraping o por compra directa en foros clandestinos[54]. La automatización permite mantener decenas de miles de contactos activos, con mensajes que se ajustan al perfil de la víctima: nombre completo, ubicación, nombres de familiares, lugar de trabajo o historial digital. Esta personalización opera como verificador de autenticidad, generando miedo y urgencia sin requerir contacto directo.

En la dimensión financiera, el modus operandi incluye el uso de criptomonedas para mover fondos producto del fraude o la extorsión. El Cártel de Sinaloa ha establecido una infraestructura de triangulación con operadores chinos que permite transferir criptoactivos desde wallets en EE.UU. o México, convertirlos en yuanes mediante corredores no regulados, y reintegrarlos en Asia como pagos a proveedores de precursores o servicios logísticos[55]. Esta dinámica elimina la necesidad de utilizar bancos o transportes físicos, reduciendo la trazabilidad del capital delictivo.

Beneficiarios y víctimas

El despliegue sistemático de tecnologías de IA por parte del CJNG y el Cártel de Sinaloa no ha representado una simple ampliación de sus capacidades tácticas, sino una transformación profunda de sus arquitecturas operativas. En este nuevo ecosistema, los principales beneficiarios no son únicamente los altos mandos tradicionales, sino también los operadores financieros, las células digitales y los especialistas en sistemas, que han logrado consolidar nodos tecnológicos de acción criminal resilientes, móviles y desmaterializados.

En ambos casos, la IA no funciona como un simple asistente técnico, sino como un multiplicador de invisibilidad operativa y de capacidad intimidatoria. Automatiza la amenaza, disuelve la exposición delictiva y amplía el espectro de impacto simbólico sin requerir confrontación física. Esta dinámica ha permitido integrar nuevos perfiles criminales: desarrolladores de software, diseñadores de deepfakes, ingenieros en datos y especialistas en blockchain, que ahora coexisten con sicarios, halcones o mulas en la economía delictiva ampliada.

Del otro lado del sistema, las víctimas son múltiples, dispersas y en su mayoría invisibilizadas. En primer lugar, los individuos en situación de vulnerabilidad digital, afectiva o económica: adultos mayores, mujeres solas, migrantes y personas con escasa alfabetización tecnológica. Estos sectores constituyen el blanco principal de campañas de fraude afectivo o extorsión familiar, construidas con información personalizada obtenida a través de scraping o bases de datos filtradas.

En segundo lugar, las microempresas, comerciantes locales y trabajadores por cuenta propia enfrentan amenazas automatizadas que simulan procesos judiciales, sanciones fiscales o denuncias falsas. Estos mensajes, enviados desde cuentas cifradas o números rotativos, reproducen identidades visuales oficiales e incluyen deepfakes de funcionarios para inducir miedo y pagos inmediatos. La desprotección institucional y la saturación de canales de denuncia amplifican la efectividad de esta coerción.

También son víctimas las comunidades locales, especialmente en regiones con débil presencia estatal o alta conflictividad. Allí, los cárteles implementan campañas de ocupación simbólica, difundiendo amenazas, comunicados falsos o rumores digitales que minan la confianza, paralizan la denuncia y normalizan la sumisión. Este tipo de coerción algorítmica no requiere control territorial directo: controla la narrativa del espacio mediante automatización del miedo.

En suma, el uso de IA por parte del CJNG y el Cártel de Sinaloa ha transformado la relación entre estructura criminal y espacio social. La violencia ya no opera únicamente desde la presencia armada, sino desde la simulación digital, la automatización del daño y la ocupación cognitiva. Como advierte el informe de UNICRI[56], el riesgo contemporáneo no reside solo en las armas, sino en los modelos algorítmicos que personalizan la amenaza, replican el discurso del miedo y desplazan la responsabilidad del perpetrador.

52 Seminario sobre Violencia y Paz. (2024, April). Criminal recruitment on TikTok: A study documents more than 100 active accounts in Mexico. El Colegio de México, Laboratorio de Odio y Concordia and Civic A.I. Lab of Northeastern University. https://violenciaypaz.colmex.mx/publicacion/nuevas-fronteras-en-el-reclutamiento-digital-estrategias-de-reclutamiento-del-crimen-organizado-en-tiktok

53 Martínez, R. (2024, August 27). This is how the CJNG uses AI to commit fraud and extortion, according to InSight Crime. Infobae. https://www.infobae.com/mexico/2024/08/27/asi-es-como-el-cnjg-utiliza-ia-para-cometer-fraudes-y-extorsiones-segun-insight-crime/

54 García, S. (2025, May 8). How criminal groups have adapted to the digital age. InSight Crime. https://insightcrime.org/es/noticias/como-grupos-criminales-adaptado-era-digital/

55 TRM Labs. (2024, July 26). Authorities unravel the Sinaloa Cartel's connection to Chinese money launderers. TRM Blog. https://www.trmlabs.com/es/resources/blog/authorities-unravel-the-sinaloa-cartels-connection-to-chinese-money-launderers

56 UNICRI. (2021). Algorithms and terrorism: The malicious use of artificial intelligence for terrorist purposes. https://unicri.org/News/Algorithms-Terrorism-Malicious-Use-Artificial-Intelligence-Terrorist-Purposes

CASO 2. ISIS (NEWS HARVEST)

El Estado Islámico (ISIS), organización terrorista internacional con estructura jerárquica rígida y una lógica doctrinaria centralizada, ha integrado de forma sofisticada tecnologías de IA para potenciar su aparato propagandístico. Desde 2023, el caso más representativo ha sido el lanzamiento del programa *News Harvest*, una serie de noticieros falsos generados con presentadores virtuales creados por IA, diseñados para amplificar la narrativa yihadista del grupo[57]. Estos videos, difundidos a través de plataformas como Telegram, Facebook, TikTok y X, han replicado el estilo visual y discursivo de noticieros, incorporando gráficos dinámicos, escenografías digitales y discursos adaptados a distintos contextos lingüísticos y culturales.

El contenido ha sido producido en varios idiomas —incluyendo árabe, inglés y urdu— con foco en países de alta vulnerabilidad ideológica o con presencia de células activas, como Irak, Siria, Afganistán, Pakistán, Indonesia, Nigeria y partes de Europa. Esta estrategia busca no solo fortalecer la cohesión interna del grupo, sino también atraer nuevos reclutas, justificar actos violentos y socavar la legitimidad de gobiernos locales. Según múltiples fuentes, como los de GNET[58], UNICRI[59] y India Times[60], ISIS ha empleado IA generativa para diseñar tanto la imagen como el discurso de sus "presentadores" sintéticos, a los que dota de apariencia confiable, tono persuasivo y fluidez en múltiples lenguas, lo que incrementa el alcance y credibilidad de sus mensajes en audiencias jóvenes y tecnológicamente conectadas.

Tecnologías utilizadas

ISIS ha hecho uso de tecnologías de clonación de voz para simular con notable fidelidad el discurso de líderes reales o ficticios, a fin de mantener la percepción de continuidad jerárquica o de presencia operativa en zonas de conflicto. Esta técnica ha sido clave para generar contenidos de autoridad, incluso cuando los líderes originales han muerto o están desaparecidos. Asimismo, se han detectado indicios del uso de reconocimiento facial inverso en materiales propagandísticos —como videos de ejecuciones, detenciones o amenazas—, los cuales permiten identificar objetivos específicos, ya sean desertores, periodistas, fuerzas armadas o disidentes religiosos, con el fin de generar intimidación dirigida y reforzar el control informativo[61].

La GenIA, tanto visual como audiovisual, ha sido fundamental en el caso del programa *News Harvest*, que utiliza presentadores sintéticos multilingües. Estos avatares imitan con alta fidelidad el formato profesional de cadenas noticiosas como Al Jazeera, BBC o CNN, incorporando fondos virtuales de estudio, lectura mediante teleprompter digital, gesticulación coordinada y lenguaje corporal simulado. Esta estética profesional refuerza la credibilidad del contenido y aumenta la capacidad de persuasión del mensaje.

Aunque no existen evidencias concluyentes sobre el uso explícito de sistemas de smart routing, diversos analistas en seguridad cibernética han sugerido que la geolocalización automatizada de audiencias, la adaptación de acentos regionales, la traducción algorítmica y la evasión de censura mediante redes distribuidas (mirror sites, VPN, enlaces efímeros) podrían estar mediadas por sistemas que optimizan la dispersión del mensaje de manera estratégica[62]. Esta capa algorítmica permitiría a ISIS adaptar el contenido propagandístico a contextos locales, reducir la exposición operativa y maximizar la efectividad de su guerra cognitiva.

Modus operandi

La estructura jerárquica de ISIS ha facilitado la integración estratégica de tecnologías de IA en tres esferas críticas de su operación: logística, coerción y propaganda. A diferencia de colectivos distribuidos o plataformas autónomas, ISIS conserva una cadena de mando clara que permite coordinar la adopción tecnológica desde sus niveles doctrinarios hasta operativos. En este marco, la IA se convierte en una herramienta de refuerzo estructural: permite optimizar la producción de contenidos, escalar la capacidad de difusión, y mantener el control doctrinal sobre la narrativa diseminada en distintos idiomas y regiones.

El contenido emitido por *News Harvest* se redacta a partir de fuentes internas del grupo, principalmente su boletín *Al-Naba*, que proporciona líneas ideológicas, posicionamientos estratégicos y actualizaciones sobre sus operaciones en zonas como Siria, Irak, Afganistán o África Occidental. Estas narrativas son posteriormente adaptadas por equipos mediáticos del grupo y transformadas en piezas audiovisuales sintéticas, las cuales son difundidas a través de plataformas encriptadas como Telegram, o replicadas masivamente en redes como Facebook, X y TikTok. Este proceso integra la GenIA para personalizar el lenguaje, simular gestos y acentos locales, y producir versiones del mismo mensaje para distintas audiencias[63].

En este contexto, la IA no solo automatiza la propaganda, sino que reconfigura el aparato comunicacional de la organización. Gracias a lo que podría denominarse una estrategia de "automatización ideológica", el grupo ha reducido sus costos operativos, eliminado la necesidad de voceros humanos —lo cual disminuye el riesgo de ubicación y neutralización—, y aumentado exponencialmente su capacidad de saturar el espacio digital con contenidos altamente persuasivos y difíciles de desmentir. Esta lógica representa una mutación en las formas tradicionales de insurgencia mediática, en las que el carisma del líder era central; ahora, es la credibilidad técnica y estética de un avatar de IA lo que sostiene la narrativa. En suma, la IA no solo sustituye funciones humanas en ISIS, sino que se ha convertido en una arquitectura simbólica clave para sostener su presencia en el terreno de la guerra cognitiva.

Beneficiarios y víctimas

Los principales beneficiarios de este modelo son los altos mandos estratégicos y operadores mediáticos del Estado Islámico, quienes han logrado consolidar una estructura propagandística resiliente, eficaz y automatizada. El uso de IA en este ámbito no solo ha incrementado la velocidad y el volumen de difusión ideológica, sino que ha fortalecido el control doctrinal sobre las narrativas emitidas, incluso en escenarios de desarticulación territorial o presión militar. La IA, en este caso, actúa como un multiplicador de cohesión organizativa y un instrumento de seducción simbólica a escala transnacional.

Las víctimas de este ecosistema mediático automatizado son múltiples y se distribuyen en varios niveles. En primer lugar, los jóvenes susceptibles a procesos de radicalización —particularmente en contextos de exclusión o violencia estructural—

constituyen el objetivo primario de los contenidos diseñados por News Harvest y otras plataformas asociadas[64]. En segundo lugar, las comunidades locales expuestas a estos discursos extremistas ven erosionadas sus dinámicas sociales por la difusión de narrativas que legitiman la violencia, la exclusión o la desinformación. Además, periodistas, defensores de derechos humanos y fuerzas del orden se convierten en blancos discursivos o simbólicos de estas campañas, muchas veces mediante suplantaciones, amenazas virtuales o manipulación de sus imágenes.

Este caso permite observar cómo una organización de estructura vertical puede no solo adaptarse al ecosistema digital, sino instrumentalizarlo de forma estratégica. En el caso de ISIS, la IA no representa una mera herramienta técnica, sino un componente estructural de su aparato de poder. La automatización de la propaganda ha hecho posible la reproducción global de un mensaje extremista sin necesidad de voceros humanos, sin límites geográficos y con una estética profesional que simula la credibilidad de medios tradicionales.

57 AIID. (2024). Incident 690: ISIS utilizes AI for propaganda videos in News Harvest program. Partnership on AI. https://incidentdatabase.ai/cite/690

58 GNET. (2024). AI-powered jihadist news broadcasts: A new trend in pro-IS propaganda production. Global Network on Extremism and Technology. https://gnet-research.org/2024/05/09/ai-powered-jihadist-news-broadcasts-a-new-trend-in-pro-is-propaganda-production/

59 UNICRI. (2021). Algorithms and terrorism: The malicious use of artificial intelligence for terrorist purposes. https://unicri.org/News/Algorithms-Terrorism-Malicious-Use-Artificial-Intelligence-Terrorist-Purposes

60 Times of India. (2024, abril 3). News Harvest: How Islamic State is using AI anchors to boost propaganda. https://timesofindia.indiatimes.com/india/news-harvest-how-islamic-state-is-using-ai-anchors-to-boost-propaganda/articleshow/110463842.cms

61 AIID. (2024). Incident 690: ISIS utilizes AI for propaganda videos in News Harvest program. Partnership on AI. https://incidentdatabase.ai/cite/690

62 UNICRI. (2021). Algorithms and terrorism: The malicious use of artificial intelligence for terrorist purposes. https://unicri.org/News/Algorithms-Terrorism-Malicious-Use-Artificial-Intelligence-Terrorist-Purposes

63 Speckhard, A., Thakkar, M. (2024, July 15). ISIS supporters harness the power of AI to ramp up propaganda on Facebook, X and TikTok. Homeland Security Today. https://www.hstoday.us/featured/is-iskp-supporters-harness-generative-ai-for-propaganda-dissemination/

64 UNICRI. (2021). Algorithms and terrorism: The malicious use of artificial intelligence for terrorist purposes. https://unicri.org/News/Algorithms-Terrorism-Malicious-Use-Artificial-Intelligence-Terrorist-Purposes

CASO 3. KK PARK

KK Park representa la consolidación de un nuevo tipo de actor criminal híbrido: una infraestructura urbana privada, financiada por mafias chinas, protegida por milicias étnicas armadas, tolerada por regímenes autoritarios y sostenida por tecnologías digitales avanzadas. Localizado en Myawaddy, en la frontera entre Myanmar y Tailandia, el complejo comenzó como un proyecto inmobiliario de lujo promovido por la empresa Yatai International Holdings, del empresario chino She Zhijiang[65]. Sin embargo, investigaciones periodísticas, testimonios de víctimas y reportes internacionales han evidenciado que KK Park opera como una ciudad-fábrica de estafas, con decenas de miles de trabajadores forzados, sometidos a tortura, esclavitud digital y trata de órganos[66].

La estructura de mando combina elementos paramilitares y empresariales. En la cúspide se encuentran redes criminales chinas como la Tríada 14K, liderada por Wan Kuok Koi ("Broken Tooth"), vinculada a la organización Hongmen y al Partido Comunista Chino. La seguridad física es controlada por milicias locales como la Karen Border Guard Force (BGF) y el Karen National Army (KNA), que actúan como ejército privado. Estas fuerzas resguardan el perímetro, controlan a los "empleados" y reprimen cualquier intento de fuga o insubordinación[67].

Al mismo tiempo, la gobernanza de KK Park opera bajo una lógica corporativo-algorítmica. Se establecen metas semanales, manuales de instrucción para el fraude, jornadas de 17 horas, mecanismos de vigilancia digital, sistemas de recompensa y castigo automatizados[68]. La coerción no es únicamente física, sino también simbólica, emocional y computacional. Este modelo trasciende el crimen organizado tradicional: se trata de una ciudad digital del crimen, donde el poder no reside en el control territorial sino en la capacidad de producción de estafas y circulación de criptomonedas.

Tecnologías utilizadas

KK Park no es solo un enclave físico de explotación humana; es una arquitectura digital del crimen que ha logrado integrar un conjunto de tecnologías de forma sistémica, al servicio de una economía de la estafa industrializada. En el corazón de su funcionamiento se encuentra la GenIA, utilizada para simular conversaciones afectivas, crear identidades falsas y diseñar contenido textual persuasivo en múltiples idiomas[69]. Estas capacidades no se limitan al engaño inicial: permiten sostener relaciones falsas durante semanas o meses, con el fin de inducir progresivamente a las víctimas a realizar inversiones en plataformas fraudulentas.

A esta capa conversacional se suma una dimensión audiovisual igualmente sofisticada. Mediante clonación de voz y deepfakes, los operadores de KK Park suplantan a familiares, autoridades o asesores financieros, generando audios y videos con un realismo inquietante. Estas herramientas son utilizadas para reforzar la ilusión de legitimidad de las operaciones, escalar emocionalmente el vínculo y aumentar la presión sobre las víctimas[70]. La ilusión es total: desde la voz que llama en tono de urgencia hasta los videos que muestran supuestos balances financieros o reuniones con supuestos ejecutivos.

El proceso de victimización se inicia, muchas veces, con técnicas de scraping masivo y minería de datos, que permiten identificar perfiles vulnerables en redes sociales, aplicaciones de citas o plataformas laborales[71]. La selección de objetivos no es aleatoria: es resultado de algoritmos que clasifican a las personas según su edad, nivel educativo, historial económico o emocional. Estas bases son luego utilizadas para alimentar bots conversacionales, entrenados para sostener largas interacciones, simular interés afectivo y conducir a las víctimas hacia dashboards falsos que imitan interfaces de inversión, comercio o trading, mostrando ganancias ficticias como anzuelo para nuevos depósitos[72].

En el núcleo financiero del sistema, KK Park opera con una lógica de lavado algorítmico sustentado en criptomonedas como Tether (USDT). Los ingresos generados por las estafas son inmediatamente transformados en activos digitales, triangulados mediante wallets anónimas, exchanges de dudosa regulación (como Binance, Huobi u OKX) y empresas fachada[73]. La capa técnica se completa con una infraestructura privada de telecomunicaciones, con antenas propias, servidores locales y conexiones encriptadas que evitan la dependencia de redes estatales. Esta soberanía tecnológica garantiza la continuidad del negocio criminal incluso frente a intentos externos de intervención.

Modus operandi

El modelo operativo de KK Park combina la eficiencia algorítmica con la brutalidad física. Miles de personas son reclutadas con falsas ofertas de empleo desde países como Kenia, Malasia, Brasil, India y Filipinas. Una vez capturadas, son trasladadas a Myanmar y retenidas en instalaciones con alambre de púas, vigilancia armada y cámaras. Allí se les obliga a trabajar en estafas digitales 17 horas al día, con castigos que van desde la privación de alimentos hasta las descargas eléctricas y el tráfico de órganos[74].

Las víctimas remotas —es decir, las personas estafadas— son identificadas mediante análisis de datos. Los operadores inician contacto afectivo por plataformas de citas, redes sociales o apps de inversión. Una vez establecida la relación emocional, se induce a la víctima a invertir en supuestas plataformas financieras. Estas páginas están diseñadas para simular ganancias iniciales, con dashboards falsos y testimonios automatizados, hasta que la víctima transfiere todo su capital[75].

A nivel interno, los trabajadores deben cumplir cuotas de rendimiento semanales. Si fallan, son amenazados con ser "vendidos" a otros centros más violentos[76]. Algunos testimonios indican que se han practicado ejecuciones extrajudiciales de quienes intentaron escapar o sabotear el sistema[77].

El lavado de dinero se ejecuta mediante una arquitectura financiera paralela basada en criptomonedas. Las ganancias se convierten en USDT, pasan por mixers o empresas pantalla, y se redistribuyen a través de cuentas controladas por actores como Wang Yi Cheng o redes vinculadas a Hongmen. Estos fondos financian, a su vez, a las milicias locales y a otros proyectos inmobiliarios en Myanmar, Laos y Camboya[78].

Beneficiarios y víctimas

Los principales beneficiarios de la operación en KK Park son actores que representan una convergencia de intereses ilícitos, militares y empresariales. En la cúspide del entramado se encuentran mafias chinas como 14K y Hongmen, cuyos líderes —como Wan Kuok Koi ("Broken Tooth")— han logrado construir una red de poder que vincula el crimen organizado, las milicias locales y el capital digital[79]. A ellos se suman empresarios con identidad oculta que son arquitectos del sistema financiero de lavado, y grupos paramilitares como la Karen Border Guard Force y el Karen National Army, que actúan como custodios del complejo[80]. Todos estos actores han encontrado en KK Park un modelo sostenible de explotación industrializada que no solo genera ganancias, sino que consolida su poder local, regional y digital.

También figuran entre los beneficiarios los intermediarios tecnológicos que diseñan y sostienen la infraestructura digital del fraude. Ingenieros en IA, especialistas en blockchain, desarrolladores de dashboards falsos y analistas de datos conforman una nueva clase de "tecnócratas del crimen", que operan desde Hong Kong, Dubái, Bangkok o incluso Europa del Este. Sus servicios permiten que la maquinaria criminal de KK Park funcione sin fricción ni interrupción. Lejos de ser actores periféricos, constituyen un eslabón clave del ecosistema.

Del otro lado se ubican las víctimas, divididas en dos niveles claramente diferenciados pero interconectados por el flujo del sufrimiento. En primer lugar están las víctimas internas, los trabajadores forzados reclutados mediante engaños, secuestrados o incluso vendidos por redes de trata. Estas personas, provenientes de África, Asia y América Latina, son despojadas de

65 PlasBit (Ziken Labs). (2024, July 7). What is KK Park Myanmar: Crypto scams and human trafficking. https://plasbit.com/blog/what-is-kk-park-myanmar

66 Head, J. (2025, February 15). Scams, casinos and skyscrapers: The luxurious ghost city that emerged in one of the world's poorest areas (and in the middle of a civil war). BBC News Mundo. https://www.bbc.com/mundo/articles/c87dq98wp58o

67 C4ADS. (2025, March 27). Hot lines: Tracing movements to and from Myanmar's scam centers. https://c4ads.org/commentary/hot-lines/

68 Bayer, J., Pineda, J., Li, Y. (2024, January 30). How Chinese mafia are running a scam factory in Myanmar. DW. https://www.dw.com/en/how-chinese-mafia-are-running-a-scam-factory-in-myanmar/a-68113480

69 Bayer, J., Sanders, L., Pineda, J., Li, Y. (2024, January 30). Human trafficking in internet scam factories. DW. https://www.dw.com/es/obligados-a-engañar-trata-de-personas-en-fábricas-de-estafas-por-internet/a-68126398

70 PlasBit (Ziken Labs). (2024, July 7). What is KK Park Myanmar: Crypto scams and human trafficking. https://plasbit.com/blog/what-is-kk-park-myanmar

71 Ziken Labs. (2024, July 7). What is KK Park Myanmar: Crypto scams and human trafficking. PlasBit. https://plasbit.com/blog/what-is-kk-park-myanmar

72 Di Girolamo, M. (2025, March 27). Hot lines: Tracing movements to and from Myanmar's scam centers. C4ADS. https://c4ads.org/commentary/hot-lines/

73 Kykyo (2024). *Chinese criminal gangs drive rise in pig-butchering scams as victims suffer emotional, financial harm* Coinlive. https://www.coinlive.com/news/chinese-criminal-gangs-drive-rise-in-pig-butchering-scams-as-victims

74 Acertpix. (2025, February 18). KK Park: The online fraud factory involved in employee exploitation. https://acertpix.com.br/blog/kk-park-a-fabrica-de-fraude-online-envolvida-em-exploracao-de-funcionarios/

75 Ziken Labs. (2024, julio 7). What Is KK Park Myanmar: Crypto Scams and Human Trafficking. PlasBit. https://plasbit.com/blog/what-is-kk-park-myanmar

76 Regan, H., Watson, I., Rebane, T., Olarn, K. (2025, April 2). Global scam industry evolving at unprecedented scale despite recent crackdown. CNN. https://edition.cnn.com/2025/04/02/asia/myanmar-scam-center-crackdown-intl-hnk-dst/index.html

77 Bayer, J., Sanders, L., Pineda, J., Li, Y. (2024, January 30). Human trafficking in internet scam factories. DW. https://www.dw.com/es/obligados-a-engañar-trata-de-personas-en-fábricas-de-estafas-por-internet/a-68126398

78 McCready, A., Mendelson, A. (2023, July 22). Myanmar: Chinese-run scam hubs reportedly continue running unabated with signs of human trafficking and forced labour. Business & Human Rights Resource Centre. https://www.business-humanrights.org/en/latest-news/myanmar-chinese-run-scam-hubs-reportedly-continue-running-unabated-with-signs-of-human-trafficking-and-forced-labour/

79 Di Girolamo, M. (2025, March 27). Hot lines: Tracing movements to and from Myanmar's scam centers. C4ADS. https://c4ads.org/commentary/hot-lines/

80 Bayer, J., Pineda, J., Li, Y. (2024, January 30). How Chinese mafia are running a scam factory in Myanmar. DW. https://www.dw.com/en/how-chinese-mafia-are-running-a-scam-factory-in-myanmar/a-68113480

sus documentos, tatuadas como mercancía, y sometidas a condiciones de esclavitud digital, donde cada interacción simulada con una víctima remota representa un acto obligado de criminalidad impuesta[81].

Se estima que desde el inicio de sus operaciones, en algún punto de 2021, más de 100,000 personas han sido víctimas de esclavitud digital en estos centros[82]. Sin embargo, se estima que actualmente KK Park alberga al menos 20,000 trabajadores-esclavos, reclutados bajo engaños laborales y obligados a operar plataformas de estafas digitales[83].

En segundo lugar se encuentran las víctimas externas, miles de personas —muchas veces ciudadanos de clase media, jubilados o migrantes— que caen en los esquemas de pig butchering, fraude amoroso o inversión. Engañadas durante semanas, inducidas emocionalmente a confiar y a comprometer sus ahorros, estas personas son despojadas de todo. Algunas nunca denuncian por vergüenza o miedo; otras terminan en bancarrota o incluso se suicidan. A este nivel, la estafa no es solo financiera: es una forma de violencia simbólica y emocional extrema, que destruye confianza, vínculos sociales y estabilidad psíquica.

A nivel institucional, las víctimas son los Estados democráticos y los sistemas financieros globales, que ven comprometida su capacidad de protección y su legitimidad ante una ola de crímenes casi invisibles, automatizados, transnacionales y sin perpetrador identificable. KK Park no representa solo un caso, sino un modelo operativo replicable, una infraestructura de criminalidad algorítmica que ser sirve de la IA y redefine el poder desde la interfaz digital. En este contexto, no basta con capturar criminales: hay que desmantelar sistemas.

Implicaciones estratégicas

La supervivencia —e incluso el fortalecimiento— de organizaciones jerárquicas tradicionales en la era algorítmica representa una paradoja inquietante: presenta un ecosistema fértil para la integración de IA como instrumento de poder, coerción y reproducción simbólica. En lugar de disolverse ante la irrupción de nuevas tecnologías, las cadenas verticales de mando se han traducido en arquitecturas digitales que permiten al crimen organizado operar con mayor eficiencia, menor exposición y una capacidad de daño profundamente ampliada.

La primera implicación estratégica es evidente: el poder jerárquico no ha desaparecido; se ha codificado. En el caso de ISIS, la verticalidad doctrinal se ha convertido en un motor de automatización ideológica, donde bots, avatares y sistemas de traducción algorítmica sustituyen al predicador físico pero mantienen intacta la autoridad del mensaje. La voz del califa ya no necesita resonar en una mezquita; basta con que un avatar sintético hable con convicción desde una pantalla para que la retórica del martirio circule globalmente. Esto representa un salto cualitativo: la centralización ideológica ha logrado sobrevivir a la descentralización geográfica, manteniendo su poder de reclutamiento, cohesión y desestabilización con costos operativos mínimos.

En paralelo, el CJNG y el Cártel de Sinaloa han demostrado que la verticalidad no sólo es compatible con la IA, sino que puede potenciarla como infraestructura de control simbólico y financiero. Lo relevante aquí no es la sofisticación técnica por sí misma, sino la capacidad de estas organizaciones para adaptar sus lógicas internas de mando —militarizada en el CJNG, celular en Sinaloa— a nuevas herramientas de suplantación, coerción emocional y automatización del miedo. La IA no funciona como gadget, sino como sistema: desde los bots conversacionales que inducen al pago bajo amenaza, hasta los algoritmos que seleccionan víctimas por perfil psicológico, lo que se observa es una integración sistémica que convierte a la verticalidad en ventaja táctica. La orden no sólo baja en la cadena de mando: se programa.

KK Park lleva esta lógica al extremo. Lo que comenzó como un centro físico de estafas se ha transformado en una ciudad-fábrica del crimen algorítmico, donde la verticalidad no se manifiesta solo en términos de poder armado, sino en la capacidad de imponer metas de rendimiento criminal, algoritmos de manipulación afectiva, y rutinas laborales coercitivas con base en software de seguimiento. Se trata de una gobernanza híbrida, en la que las milicias, el capital criminal y la IA coexisten en un régimen autoritario de explotación digital.

Este tipo de configuraciones obliga a repensar el concepto mismo de “poder criminal”. Ya no se trata únicamente de controlar territorios o ejercer violencia armada, sino de construir infraestructuras simbólicas, emocionales y financieras que permiten ejercer dominio a distancia. El liderazgo se vuelve modular, las operaciones se disocian del cuerpo, y la amenaza se automatiza. En este contexto, la verticalidad se vuelve invisible pero no menos efectiva: no necesita presencia física, sino persistencia digital.

Para los Estados, esto representa una disrupción estructural. Las herramientas tradicionales de seguridad —captura de líderes, decomisos, interdicciones— pierden eficacia frente a organizaciones que pueden replicar su mando mediante avatares, mover dinero sin tocar billetes y extorsionar sin levantar el teléfono. La lógica de la persecución judicial se ve superada por una realidad donde el perpetrador puede ser una interfaz, la amenaza un algoritmo, y el botín una línea de código.

Los sistemas legales enfrentan una arquitectura del crimen que no puede ser desmantelada con estrategias del siglo XX. Se requiere una reconceptualización urgente del marco normativo, técnico y diplomático que permita enfrentar a estos actores como lo que realmente son: sistemas híbridos de gobierno algorítmico con capacidad transnacional.

Además, el fenómeno tiene una dimensión geopolítica. ISIS, CJNG y KK Park no operan en vacío: su existencia está mediada por contextos de colapso institucional, protección estatal informal o complicidad activa de actores estatales y paraestatales. La exportación del daño, la fragmentación de responsabilidades y la opacidad operativa hacen que las respuestas unilaterales sean ineficaces. La naturaleza transnacional, digital y estructurada de estas organizaciones demanda una coordinación internacional sin precedentes, no sólo entre agencias de seguridad, sino entre organismos regulatorios, actores tecnológicos, instituciones financieras y entidades de gobernanza digital. La contención no se logrará con fuerza, sino con interoperabilidad estratégica.

Las organizaciones jerárquicas tradicionales, lejos de extinguirse, han mutado en arquitecturas de poder programado. Su persistencia no es un residuo del pasado, sino una advertencia del futuro: allí donde el mando encuentra en la IA un aliado, el crimen no desaparece. Se vuelve invisible, eficiente, replicable. El verdadero desafío ya no es capturar al jefe: es apagar el sistema.

81 Head, J. (2025, February 15). Scams, casinos and skyscrapers: The luxurious ghost city that emerged in one of the world’s poorest areas (and in the middle of a civil war). BBC News Mundo. https://www.bbc.com/mundo/articles/c87dq98wp58o

82 Regan, H., Watson, I., Rebane, T., Olarn, K. (2025, April 2). Global scam industry evolving at unprecedented scale despite recent crackdown. CNN. https://edition.cnn.com/2025/04/02/asia/myanmar-scam-center-crackdown-intl-hnk-dst/index.html

83 Ziken Labs. (2024, July 7). What is KK Park Myanmar: Crypto scams and human trafficking. PlasBit. https://plasbit.com/blog/what-is-kk-park-myanmar

REDES DISTRIBUIDAS O CIBERCOLECTIVOS

A diferencia de las organizaciones jerárquicas tradicionales, cuyo poder se articula a través de cadenas de mando, coerción física y control territorial, las redes criminales distribuidas operan bajo una lógica radicalmente distinta: informalidad estructural, autonomía nodal y simbiosis algorítmica. No obedecen a una figura carismática, no controlan barrios ni necesitan bases territoriales; su dominio se ejerce desde el anonimato, en servidores federados, canales efímeros y plataformas de acceso cifrado. Lejos de ser marginales, estas redes representan hoy uno de los desafíos más complejos para la seguridad internacional: el crimen sin jerarquía visible, sin geografía estable y sin rostro humano.

Estas organizaciones—o más precisamente, estos ecosistemas delictivos—nacen en la confluencia del ciberactivismo, el oportunismo criminal y la economía de servicios ilegales. Su morfología no es piramidal ni militar, sino reticular y adaptativa: nodos semiautónomos que cooperan mediante herramientas compartidas, afinidades funcionales y lógicas de oportunidad. Algunas se originan en el fraude digital—como los Yahoo Boys nigerianos—, otras en el ciberactivismo que derivó hacia operaciones delictivas, como FunkSec, y otras más emergen de contextos de exclusión estructural y conectividad fragmentada, como el Sindicato del Piso 13 de Poipet o los colectivos “montadeudas” en Ciudad de México. A pesar de su diversidad geográfica y cultural, todas comparten una característica: la IA no es un apoyo técnico, sino el catalizador de su existencia operativa.

La IA, en este contexto, opera como pegamento técnico, multiplicador de capacidad y acelerador de escala. Herramientas de scraping masivo, generación de identidades sintéticas, automatización conversacional, bots de targeting y clonadores de voz reemplazan las funciones humanas sin comprometer la eficacia. Esta externalización algorítmica no solo permite el funcionamiento distribuido: lo vuelve inevitable. La descentralización, antes sinónimo de debilidad táctica, se ha convertido en un blindaje operativo.

Estas redes operan bajo lógicas colaborativas propias del ecosistema digital: comparten scripts, adaptan plantillas de estafa, intercambian listas de víctimas, actualizan modelos de lenguaje para fraudes localizados y utilizan foros como Exploit.in, breach forums o canales alternativos de Telegram para resolver disputas internas y refinar estrategias. A diferencia de los carteles o las mafias, su poder no depende de la obediencia, sino de la interoperabilidad. No construyen lealtad: construyen compatibilidad funcional.

Pero su informalidad no implica fragilidad. Al contrario: su flexibilidad estructural las hace especialmente resistentes a la desarticulación. Cuando un nodo es detectado o eliminado, otros lo reemplazan de forma casi instantánea. La disolución es parte de su diseño. Estas redes no mueren: se actualizan. Cada actor es reemplazable, pero la arquitectura persiste, impulsada por herramientas open source, APIs públicas y un mercado creciente de servicios algorítmicos que permiten que la criminalidad fluya, sin necesidad de liderazgo ni ideología. Solo eficiencia y lucro.

En los casos que siguen—FunkSec, Yahoo Boys, Montadeudas, Poipet y Operación Cumberland—se observa con claridad esta lógica distribuida y esta economía de la criminalidad sin centro. Ninguno de estos grupos controla un territorio. Ninguno tiene un líder carismático. Ninguno mantiene una narrativa doctrinal. Y sin embargo, todos han demostrado una capacidad devastadora para ejecutar fraudes masivos, extorsión automatizada, producción de CSAM sintético y sabotajes digitales dirigidos. Son colectivos, cooperativas o enjambres que funcionan como interfaces del crimen algorítmico, y cuya estructura fluida representa una amenaza estructural para la trazabilidad, la persecución penal y la cooperación internacional. En el siglo XXI, el crimen ya no necesita levantar una bandera ni controlar una esquina: le basta con ejecutar un script y desaparecer.

CASO 1. FUNKSEC

FunkSec no es simplemente un grupo emergente de ransomware; es el síntoma más inquietante de una transformación epistémica en el crimen organizado digital: el paso de la jerarquía a la diseminación, del poder armado al poder algorítmico, del comando físico a la ejecución simbólica descentralizada[84].

Desde su aparición pública en diciembre de 2024, FunkSec se posicionó con una velocidad inusitada como uno de los actores más prolíficos del ecosistema ransomware-as-a-service (RaaS). En su primer mes de operación superó las 120 víctimas, con campañas activas en al menos 47 países[85]. Su ascenso no fue impulsado por sofisticación técnica sino por una arquitectura modular que permite a operadores inexpertos lanzar campañas de ransomware funcionales con apoyo de asistentes de GenIA. Esta economía criminal de código y afiliación representa el paso del crimen organizado al crimen distribuido.

El grupo se autodefine como pro-Palestina y antiimperialista. Sus comunicados, sus ransom notes y su presencia en foros como Exploit.in o Telegram replican una estética hacktivista que remite a Ghost Algeria o Cyb3r Fl00d. Sin embargo, su operación es típicamente extorsiva, con campañas de doble amenaza: cifrado de archivos y filtración pública en su Data Leak Site, seguido de subastas en la plataforma FunkBID[86]. Esta ambigüedad entre activismo digital y crimen financiero constituye uno de los desafíos más urgentes para los marcos regulatorios actuales, incapaces aún de diferenciar la simulación ideológica del cibercrimen motivado exclusivamente por rentabilidad.

Tecnologías utilizadas

La IA cumple en FunkSec un papel que excede el soporte técnico. Es la arquitectura misma de su funcionamiento. Modelos de lenguaje como GPT-4, Claude o Miniapps han sido instrumentalizados para generar scripts de cifrado, redactar mensajes extorsivos en inglés técnico y simular interfaces de soporte. El ransomware está escrito en Rust, con capas redundantes que dificultan la ingeniería inversa y cifrado híbrido RSA–AES que ha desafiado incluso a antivirus comerciales[87].

84 AIID. (2025). Incident 897: AI-assisted ransomware campaign by FunkSec allegedly targets over 80 victims. https://incidentdatabase.ai/cite/897

85 SOCRadar. (2025, January 4). Dark web profile: FunkSec. SOCRadar Cyber Intelligence Inc. https://socradar.io/dark-web-profile-funksec/

86 FireXCore. (2025, May 25). AI-driven ransomware FunkSec: The shocking fusion of hacktivism and cybercrime. https://firexcore.com/blog/ai-driven-ransomware-funksec/

87 Check Point Software. (2025, May). FunkSec ransomware – AI powered group. https://www.checkpoint.com/cyber-hub/threat-prevention/ransomware/funksec-ransomware-ai-powered-group/

Para su operación, FunkSec desarrolló un bot de control remoto JQRAXY_HVNC que permite acceso encubierto en sistemas infectados con evasión de logs. Además, el colectivo ha lanzado generadores automáticos de credenciales, herramientas de DDoS y entornos de prueba para campañas de spear phishing en múltiples idiomas[88]. FunkSec ha desarrollado una infraestructura criminal compuesta por una plataforma de filtraciones (DLS), una subasta de datos robados (FunkBID), y módulos ofensivos que se actualizan de forma iterativa gracias a la asistencia de modelos generativos[89]. De esta forma, la IA no solo facilita el crimen: lo produce, lo disemina y lo estandariza.

Modus operandi

La lógica operativa de FunkSec no responde a campañas de alto valor. Su meta no es interrumpir infraestructuras críticas, sino masificar extorsiones de bajo monto. En ese sentido, sus rescates promedio rondan los $10,000 USD, cifra diseñada para maximizar el número de pagos sin activar respuestas estatales de alta intensidad[90]. Esta economía del daño menor, sostenida por IA y diseminada por afiliados, convierte a FunkSec en una red de extorsión de bajo umbral y alta persistencia.

El colectivo ha mantenido alianzas funcionales con actores como FSociety y posiblemente con variantes del grupo Babuk. En enero de 2025 anunciaron su transición al modelo “manada” (wolf pack): campañas conjuntas, código compartido, servicios de Flocker ransomware y soporte 24/7 mediante IA[91].

88 Bitdefender Enterprise. (2025, March 4). FunkSec: An AI-centric and affiliate-powered ransomware group. https://www.bitdefender.com/en-us/blog/businessinsights/funksec-an-ai-centric-and-affiliate-powered-ransomware-group

89 SOCRadar. (2025, January 4). Dark web profile: FunkSec. SOCRadar Cyber Intelligence Inc. https://socradar.io/dark-web-profile-funksec/

90 Cyber Florida at University of South Florida. (2025, January 29). FunkSec: A top ransomware group leveraging AI. https://cyberflorida.org/funksec-a-top-ransomware-group-leveraging-ai/

91 Bitdefender Enterprise. (2025, March 4). FunkSec: An AI-centric and affiliate-powered ransomware group. https://www.bitdefender.com/en-us/blog/businessinsights/funksec-an-ai-centric-and-affiliate-powered-ransomware-group

En consecuencia, FunkSec no es ya de una banda de ciberdelincuentes, sino de una federación simbólica que intercambia tácticas, infraestructura y públicos, articulada por una lógica de mercado y no por ideología. Esta transición de la jerarquía al mercado criminal se ve acentuada por los mecanismos internos de reputación, puntuación y porcentaje de rescate. En lugar de lealtad organizacional, existe incentivo financiero.

Beneficiarios y víctimas

Las víctimas de FunkSec son dispersas, pero reveladoras. Universidades en Brasil, hospitales en India, municipios en Colombia, entidades financieras en Mongolia. El patrón es claro: estructuras públicas o mixtas con alta dependencia digital y escasa capacidad defensiva. Vectra AI señala que al menos 30% de los ataques se basan en datos reciclados, reutilizados o comprados en foros clandestinos, lo cual indica una lógica de revictimización que extiende el daño en el tiempo y lo convierte en una práctica de desgaste institucional[92].

La información robada —historiales médicos, expedientes académicos, contratos confidenciales— no solo es vendida, sino también empleada para nuevas campañas, simulaciones de identidad y fraudes escalonados. La víctima no es un blanco, sino un recurso regenerativo. Los beneficiarios inmediatos son los afiliados del colectivo, quienes reciben hasta el 70% de los pagos de rescate, acceso a herramientas automatizadas y foros de soporte IA-asistido[93]. Pero el verdadero beneficiario es el modelo: FunkSec como marca, como sistema replicable y nueva forma de operar.

FunkSec encarna, en suma, una forma de criminalidad posthumana. No por su tecnología —que no es nueva—, sino por su estructura. Es una organización sin cuerpo, sin historia y sin centralidad. Es una interfaz criminal basada en IA, legitimada por foros, reforzada por narrativas políticas y sostenida por automatismos simbólicos. Su análisis exige una reconsideración del sujeto criminal y una cartografía de los nuevos vectores de poder que se han desanclado del cuerpo, del territorio y de la historia. En el modelo FunkSec, la IA no asiste al crimen: lo reemplaza.

92 Vectra AI. (2025, May). Is your organization safe from FunkSec? https://www.vectra.ai/threat-hunting/threat-actors/funksec

93 FireXCore. (2025, May 25). AI-driven ransomware FunkSec: The shocking fusion of hacktivism and cybercrime. https://firexcore.com/blog/ai-driven-ransomware-funksec/

CASO 2. CLAN SAN ROQUE (BOLIVIA)

En 2025, una red criminal desde el penal de San Roque en Chuquisaca, Bolivia, empleó IA para clonar la voz del ministro de Educación, Omar Véliz Ramos. Este grupo, al que se denominó Clan San Roque, operó desde el interior del penal en un esquema de coordinación de presidiarios con cómplices externos, creando un sofisticado sistema de estafa digital que implicaba el uso de GenIA[94]. La estructura combinaba el poder informal carcelario con un modelo de simulación institucional digital, reproduciendo con precisión los discursos, la entonación y los protocolos de comunicación de las autoridades del Estado[95].

Lejos de ser una operación rudimentaria, la red contaba con equipos dentro y fuera del penal: captadores digitales en redes sociales, mulas financieras reclutadas en situación de calle, operadores de mensajería automatizada y diseñadores de contenido persuasivo en plataformas como TikTok, Facebook e Instagram[96]. Esta convergencia entre crimen tradicional y crimen algorítmico revela la configuración de un actor criminal híbrido que opera desde los márgenes físicos del Estado, pero con alta capacidad de penetración simbólica.

Tecnologías utilizadas

El eje operativo de la estafa fue la voz del ministro Veliz Ramos, clonada mediante tecnologías de IA entrenadas para replicar timbre, acento y ritmo con notable fidelidad. A través de esta voz sintética, el Clan San Roque realizaba llamadas personalizadas a víctimas previamente seleccionadas, ofreciéndoles falsos puestos de trabajo —conocidos como "ítems"— en dependencias públicas del Estado. Las víctimas eran inducidas a realizar pagos que oscilaban entre Bs 3,500 y Bs 5,000, habitualmente mediante códigos QR generados desde cuentas de terceros[97].

Este proceso se apoyaba también en sistemas de respuesta automática por mensajería instantánea y en la manipulación de algoritmos de recomendación para posicionar sus contenidos fraudulentos en redes sociales[98]. La segmentación emocional de las víctimas —jóvenes desempleados, mujeres jefas de hogar, adultos mayores con experiencia docente— fue facilitada por el uso de herramientas de geolocalización y análisis de perfiles en línea.

Modus operandi

La operación del Clan San Roque puede reconstruirse como un ciclo cerrado de manipulación institucional algorítmica. Todo iniciaba con publicaciones en redes sociales que simulaban ser convocatorias del Ministerio de Educación, con imágenes institucionales y lenguaje técnico verosímil[99]. Una vez que las víctimas mostraban interés, eran contactadas directamente por WhatsApp o por llamada telefónica. En estos contactos se utilizaba la voz sintética del ministro para establecer una comunicación formal y creíble, en la que se explicaban los supuestos requisitos administrativos.

Se solicitaba entonces el pago de un monto por "gastos de trámite", utilizando códigos QR vinculados a cuentas bancarias de mulas digitales. Estas cuentas eran abiertas por personas captadas mediante pagos o engaños, frecuentemente en situación de calle o extrema precariedad[100]. Tras recibir el dinero, los operadores externos desactivaban los perfiles digitales, eliminaban cuentas y bloqueaban números, haciendo muy difícil el rastreo. Todo el proceso estaba diseñado para simular legitimidad, producir confianza y luego desaparecer digitalmente sin dejar rastro para las autoridades.

Beneficiarios y víctimas

Los principales beneficiarios fueron los internos del penal de San Roque y sus cómplices externos, quienes lograron defraudar al menos a 19 personas, obteniendo ganancias superiores a los Bs 5 millones[101]. Aunque la cifra puede parecer modesta en el contexto del crimen transnacional, resulta

94 Ministerio de Educación de Bolivia. (2025, February 10). Criminal organization used artificial intelligence to clone the voice of the Minister of Education, Omar Véliz Ramos. https://www.minedu.gob.bo/index.php?option=com_content&view=article&id=7887

95 AIID. (2025). Incident 937: Bolivian criminal network uses AI voice clone of education minister. https://incidentdatabase.ai/cite/937

96 Alvarado Flores, M.E. (2025, February 10). Criminal organization used artificial intelligence to simulate the voice of the Minister of Education and commit fraud. Visión 360. https://www.vision360.bo/noticias/2025/02/10/19886-organizacion-criminal-utilizo-inteligencia-artificial-para-simular-la-voz-del-ministro-de-educacion-y-cometer-estafas

97 El Deber. (2025, February 10). Criminal organization dismantled after using the voice of the Minister of Education to defraud. https://eldeber.com.bo/pais/desbaratan-organizacion-criminal-que-usaba-la-voz-del-ministro-de-educacion-para-estafar_503161

98 Bolivia Verifica. (2025). Artificial intelligence is used to defraud using deepfakes. https://www.tiktok.com/@boliviaverifica/video/7224849569316654342

99 Ministerio de Educación de Bolivia. (2025, February 10). Criminal organization used artificial intelligence to clone the voice of the Minister of Education, Omar Véliz Ramos. https://www.minedu.gob.bo/index.php?option=com_content&view=article&id=7887

100 Alvarado Flores, M.E. (2025, February 10). Criminal organization used artificial intelligence to simulate the voice of the Minister of Education and commit fraud. Visión 360. https://www.vision360.bo/noticias/2025/02/10/19886-organizacion-criminal-utilizo-inteligencia-artificial-para-simular-la-voz-del-ministro-de-educacion-y-cometer-estafas

101 Agencia Boliviana de Información. (2025, February 10). Criminal organization cloned Minister Véliz's voice with AI, defrauded 19 people by selling positions and obtained over Bs 5 million. https://www.abi.bo/index.php/component/content/article/38-notas/noticias/seguridad/60457

significativa en tanto fue ejecutada desde una prisión y con una arquitectura tecnológica distribuida. Cada actor cumplía una función específica: unos producían contenido audiovisual, otros gestionaban las cuentas, y otros, simplemente, prestaban su identidad.

Las víctimas, en su mayoría ciudadanos desempleados o subempleados con aspiraciones legítimas de ingresar al servicio público, sufrieron no solo una pérdida económica, sino una desestabilización emocional y simbólica. La esperanza depositada en una fuente oficial de empleo fue brutalmente traicionada, generando efectos prolongados: ansiedad, miedo, vergüenza, y en algunos casos, aislamiento social. El crimen no apuntó al patrimonio como único objetivo, sino a la confianza como mecanismo de desposesión.

Este caso representa un punto de quiebre en la relación entre tecnología, penalidad y gobernanza institucional. Que una voz ministerial pueda ser clonada y usada desde una cárcel para estafar ciudadanos en nombre del Estado no solo vulnera la seguridad digital, sino que erosiona el núcleo mismo de la legitimidad democrática[102]. La estafa no fue solo económica: fue simbólica. Erosiona la credibilidad de las instituciones, y pone en duda la autenticidad de toda comunicación gubernamental.

Desde una perspectiva de gobernanza algorítmica, este caso obliga a pensar en tres líneas estratégicas: la verificación biométrica de comunicaciones institucionales, el desarrollo de sistemas antifraude contra deepfakes, y la consolidación de unidades de ciberinteligencia penitenciaria. Las prisiones ya no son un espacio aislado del crimen digital, sino uno de sus centros operativos.

102 AIID. (2025). Incident 937: Bolivian criminal network uses AI voice clone of education minister. https://incidentdatabase.ai/cite/937

CASO 3. BANDAS DE "MONTADEUDAS" CDMX (MÉXICO)

Desde 2020, la Ciudad de México ha sido epicentro de una de las modalidades más sofisticadas de extorsión algorítmica en América Latina: las bandas "montadeudas". Aunque inicialmente descritas como esquemas de préstamos irregulares, estas organizaciones operaban a través de estructuras criminales de gobernanza digital, con una arquitectura modular que integra desarrollo de software, scraping masivo de datos, coerción psicológica automatizada y manipulación simbólica de la legalidad financiera[103].

A diferencia de organizaciones verticales tradicionales, las bandas montadeudas operan como plataformas criminales distribuidas. Su estructura se basa en una red de aplicaciones móviles que cambian de nombre, logotipo y dominio constantemente, evadiendo los controles de las tiendas digitales y adaptándose rápidamente a las medidas regulatorias. Este cambio constante en la identidad digital —similar al morphing aplicado en contenido gráfico— revela su naturaleza mimética y transnacional.

La respuesta de las autoridades mexicanas a las bandas de "montadeudas" ha revelado no solo el alcance de estas redes, sino también las limitaciones estructurales del aparato jurídico frente al crimen digital replicativo. Desde mediados de 2022, la Unidad de Inteligencia Financiera (UIF), en colaboración con la Policía Cibernética y la Fiscalía General de Justicia de la Ciudad de México, documentó un incremento del 454% en las operaciones de aplicaciones fraudulentas, que operaban desde centros de llamadas disfrazados de oficinas legales, ubicados en colonias céntricas como la Juárez y Doctores[104].

En un operativo de alto perfil, se incautaron más de 700 celulares, 15,000 chips, 400 computadoras y millones de datos personales, evidencia que confirma que no se trata de estafas aisladas, sino de una infraestructura de extracción algorítmica profesionalizada. Derivado de ello, se congelaron cuentas bancarias y se bloquearon a 29 personas físicas y morales del sistema financiero, con 35 denuncias penales por extorsión, fraude, suplantación de identidad y uso ilegal de datos biométricos[105].

La dimensión transnacional del fenómeno quedó al descubierto con el rastreo de flujos financieros que conectaban a estas redes con empresas radicadas en Hong Kong, China, Costa Rica y Colombia, incluyendo depósitos en cuentas fachada por más de 70 millones de pesos mexicanos, y movimientos mensuales de hasta 219,000 dólares en transferencias digitales. Asimismo, se identificaron más de 1,046 aplicaciones activas asociadas al modelo de extorsión montadeudas, de las cuales se desactivaron 556, aunque muchas reaparecieron bajo nuevos nombres y dominios a los pocos días[106].

El proceso judicial de los detenidos, sin embargo, ha dejado grietas: en noviembre de 2023, 23 de los implicados en el principal operativo fueron liberados para enfrentar el proceso en libertad, pese a que la evidencia incluía pruebas de su participación directa en la gestión de datos personales y distribución de amenazas automatizadas. En este contexto, el Estado persigue sombras digitales que mutan más rápido que las leyes, enfrentando una forma de criminalidad que, lejos de ocultarse, se presenta con íconos de "ayuda financiera".

Tecnologías utilizadas

El corazón tecnológico del esquema montadeudas se basa en la minería de datos y el scraping automatizado. Las aplicaciones se presentan como servicios de microcrédito inmediato; sin embargo, su objetivo real es la captura masiva de datos personales, desde fotografías y contactos hasta mensajes y geolocalizaciones[107]. Estos datos se procesan para construir perfiles coercitivos hiperpersonalizados.

Una segunda capa tecnológica incluye bots de amenaza, que simulan funcionarios públicos o abogados para intimidar a las víctimas. Estos bots automatizados utilizan lenguaje jurídico simulado, referencias a códigos penales y amenazas explícitas de embargo, boletinamiento y arresto. Además, se ha documentado el uso de tecnología de morphing y generación de deepfakes, empleada para fabricar imágenes falsas de la víctima en situaciones comprometedoras —desnudez, consumo de drogas, actividades ilegales— utilizadas como medio de extorsión emocional y social. Las plataformas también implementan infraestructura elusiva, como servidores alojados en China y triangulaciones financieras en criptomonedas, lo que complica su rastreo y judicialización[108].

103 Consejo Ciudadano para la Seguridad y Justicia CDMX. (2022). Montadeudas typology: Analysis and recommendations. https://www.gob.mx/cms/uploads/attachment/file/873271/Tipolog_a_montadeudas_VF.PDF

104 López Ponce, J. (2025, January 27). How digital predatory loan scams operate in Mexico: UIF combats psychological extortion Black Mirror style. Milenio. https://www.milenio.com/policia/como-operan-los-montadeudas-digitales-en-mexico-uif

105 Martínez A. (2023, June 26). Debt app detainees avoid pretrial detention. Milenio: https://www.milenio.com/policia/detenidos-por-montadeudas-libran-prision-preventiva

106 Publimetro. (2024, March 11). Predatory loan scams: List of fraudulent apps dismantled in Mexico City. https://www.publimetro.com.mx/noticias/2022/08/18/montadeudas-lista-de-apps-fraudulentas-que-desmantelaron-en-cdmx/

107 ADN40. (2024). Predatory loan apps in Mexico 2024: Complete list and how to avoid scams. https://www.adn40.mx/mexico/apps-montadeudas-en-mexico-2024-lista-completa-actualizaa-como-evitar-las-estafas

108 Secretaría de Hacienda y Crédito Público. (2024). National risk assessment on money laundering and terrorist financing. https://www.finanzaspublicas.hacienda.gob.mx/work/models/Finanzas_Publicas/docs/congreso/infotrim/2024/it/04afp/itanfp11_202401.pdf

Modus operandi

El esquema de operación se articula en cinco fases. 1) Todo comienza con la captación, mediante anuncios en redes sociales o motores de búsqueda que prometen créditos sin buró a través de apps de préstamo. 2) Al instalar la app, la víctima autoriza el acceso total a su dispositivo, lo que permite la extracción automática de información privada[109].

3) Enseguida se activa el ciclo de intimidación, que inicia con notificaciones de una supuesta deuda impagada, usualmente inexistente, pero sujeta a intereses ficticios que crecen diariamente hasta alcanzar niveles impagables. 4) Luego, mediante llamadas, mensajes y bots, comienza una extorsión escalonada: amenazas, chantaje emocional, difamación y uso de imágenes manipuladas. En muchos casos, la humillación pública es el mecanismo de obediencia más efectivo[110]. 5) Finalmente, si la víctima cede, se le ofrece "otro préstamo" para liberarse del primero, perpetuando el ciclo de chantaje y dependencia financiera.

109 Dueñas, D. (2023, June 26). How to avoid predatory loan scams. Capital 21. https://www.capital21.cdmx.gob.mx/noticias/?p=43214

110 Infobae. (2023). What are predatory loan scams and how do they operate? Infobae México. https://www.infobae.com/mexico/2023/06/12/que-son-y-como-operan-los-montadeudas/

Víctimas y beneficiarios

En el corazón de este modelo criminal se encuentra una economía de la extorsión algorítmica que ha desplazado al crédito legítimo. Los principales beneficiarios no son prestamistas ni instituciones financieras, sino desarrolladores de software ilícito, programadores de bots coercitivos, brokers de bases de datos filtradas y operadores que actúan como mulas digitales, canalizando pagos por vías cripto o bancarias fragmentadas para evadir el rastreo tradicional. A diferencia del modelo bancario clásico, donde la ganancia emerge del interés compuesto, aquí la utilidad se concentra en la extorsión inmediata, personalizada y sin garantías: una economía criminal que monetiza el miedo.

La infraestructura criminal que sustenta a los montadeudas actúa como plataforma de captura social: datos personales, fotos íntimas, redes de contacto, huellas biométricas y hasta gestos faciales son transformados en activos de coacción. El resultado es una economía simbólica de vigilancia y humillación, donde el control emocional se convierte en producto comercializable.

Del otro lado, las víctimas no solo son sujetos vulnerables, sino blancos algorítmicamente seleccionados. En su mayoría se trata de mujeres jóvenes, jefas de hogar, trabajadoras del sector informal o empleadas domésticas, lo que demuestra una clara feminización del riesgo digital[111]. Más del 58% de las denuncias provienen de mujeres, muchas de ellas revictimizadas por la exposición de sus imágenes manipuladas o por amenazas de notificación a sus entornos laborales y familiares.

Las consecuencias trascienden lo financiero: se documentan casos de aislamiento social, despidos laborales, crisis familiares, tentativas de suicidio y pérdida total de confianza en las instituciones públicas. La violencia ejercida no depende del contacto físico, sino del dominio psicológico automatizado y la escenificación pública del castigo.

111 Consejo Ciudadano para la Seguridad y Justicia CDMX. (2022). Montadeudas typology: Analysis and recommendations. https://www.gob.mx/cms/uploads/attachment/file/873271/Tipolog_a_montadeudas_VF.PDF

CASO 4. YAHOO BOYS (NIGERIA)

A diferencia de los cárteles jerárquicos o las redes insurgentes con estructura vertical, los Yahoo Boys representan un modelo informal, distribuido y transnacional de cibercriminalidad emergente. Su configuración no responde a una cadena de mando tradicional, sino a un ecosistema de aprendizajes entre pares, círculos de iniciación conocidos como HK (*Hustling Knowledge*) y una lógica de mentoría criminal digital[112]. Este ecosistema criminal, nacido en Nigeria a partir de las estafas tipo 419 en los años noventa, ha mutado en una red expansiva que opera desde cibercafés, casas compartidas o dispositivos móviles, sin necesidad de territorio físico ni rostro definido[113]. Lo que distingue a los Yahoo Boys no es tanto la sofisticación técnica como su capacidad para transformar las emociones humanas —afecto, deseo, culpa o esperanza— en vectores de coerción.

La naturaleza distribuida de los Yahoo Boys no responde al vacío de liderazgo, sino a una lógica distinta de cohesión criminal: una pedagogía informal, en la que el conocimiento se transmite no por jerarquía, sino por imitación, reputación y acceso a recursos digitales. A diferencia de las organizaciones mafiosas tradicionales, en las que el ingreso está mediado por rituales de violencia, aquí el capital simbólico radica en la destreza para manipular emocionalmente, en la habilidad para construir una narrativa persuasiva y en la capacidad de sostenerla algorítmicamente a lo largo del tiempo.

El cibercrimen, en este contexto, no se hereda: se aprende, se practica y se optimiza colectivamente. Los Yahoo Boys son, en efecto, el resultado de una fusión entre precariedad estructural, creatividad digital y aspiraciones de movilidad social criminal, moldeadas en entornos de exclusión y desigualdad que encuentran en la estafa romántica no solo un negocio, sino una estrategia de sobrevivencia y afirmación[114].

Tecnologías utilizadas

Su transición digital se consolidó con el acceso masivo a internet, la caída en los precios de los smartphones y la disponibilidad de herramientas de IA accesibles. Lo que comenzó como ingeniería social rudimentaria mediante correos electrónicos falsos ha derivado en un ecosistema criminal tecnológicamente articulado[115]. Este se vale de algoritmos generativos, bots conversacionales, tecnologías de clonación de voz y video, y simulaciones audiovisuales capaces de engañar a múltiples víctimas en simultáneo. La IA, en este contexto, no solo cumple una función instrumental; es la arquitectura que permite simular afecto, generar falsa confianza, manipular emocionalmente y suplantar digitalmente la identidad con una eficiencia impensable hace apenas una década.

112 Ojedokun, U.A., Ilori, A.A. (2021). Tools, techniques and underground networks of Yahoo-boys in Ibadan City, Nigeria. International Journal of Criminal Justice 3, 99–122. https://doi.org/10.36889/IJCJ.2021.003

113 Barragán, C. (2023, July 11). Inside the world of Nigerian Yahoo boys. Longreads / The Atavist Magazine. https://longreads.com/2023/07/11/inside-the-world-of-nigerian-yahoo-boys-atavist-excerpt/

114 Oloworekende, A. (2019, August 28). Yahoo Yahoo – Nigeria and cybercrime's global ecosystem. The Republic. https://republic.com.ng/library/yahoo-yahoo-naija/

115 Caulfield, J. (2024). The Yahoo-boys and the upsurge in sextortion – Part 1 & 2. Linkedin. https://www.linkedin.com/pulse/yahoo-boys-upsurge-sextortion-part-1-john-caulfield-5avke

La apropiación tecnológica por parte de los Yahoo Boys ha sido especialmente significativa en el campo de la suplantación de identidad emocional. Casos como el uso de tecnología deepfake para simular la voz e imagen de Brad Pitt y estafar a una mujer francesa por más de 850,000 dólares[116], o la implementación de avatares sintéticos en videollamadas románticas en tiempo real para seducir víctimas vulnerables[117], muestran el grado de madurez que ha alcanzado este tipo de crimen distribuido. A esto se suma el uso de videos generados por IA que imitan noticieros como CNN o Fox News, donde la víctima aparece falsamente acusada de delitos sexuales con el fin de extorsionarla mediante el miedo, la vergüenza o la culpa[118]. Estas simulaciones se producen en minutos con aplicaciones comerciales, pero su impacto psicológico puede ser devastador y prolongado.

Otro elemento fundamental en su estrategia tecnológica es la creación de identidades falsas mediante generadores de rostros sintéticos, combinados con scripting emocional diseñado para sostener conversaciones románticas durante semanas. Estas conversaciones, en muchos casos, son llevadas por bots capaces de ajustar su tono, contenido y emocionalidad en función de la interacción de la víctima[119]. La capacidad de los Yahoo Boys para adaptar estas herramientas a contextos locales, lingüísticos y culturales mediante traducción automática y geolocalización refuerza su penetración global.

Modus operandi

El modus operandi de los Yahoo Boys se estructura en cuatro fases[120]: 1) La primera es la captura emocional (bombing). El contacto inicial se realiza generalmente por redes sociales o plataformas de citas. Utilizando perfiles falsos de soldados, médicos o viudos, se establece un vínculo afectivo sostenido a través de atención constante, narrativa trágica o heroica, y símbolos de confianza como fotografías familiares o mensajes en video. En esta etapa se construye una dependencia emocional que desplaza el juicio racional de la víctima.

116 AIID. (2025). Incident 901: Yahoo boys allegedly used deepfake technology to impersonate Brad Pitt and defraud French woman of $850,000 in romance scam. https://incidentdatabase.ai/cite/901

117 AIID. (2025). Incident 911: Yahoo boys allegedly employ real-time deepfake technology in romance scams. https://incidentdatabase.ai/cite/911

118 AIID (2025). Incident 913: Yahoo boys allegedly using AI-generated news videos to blackmail sextortion victims. https://incidentdatabase.ai/cite/913

119 Ojedokun, U.A., Ilori, A.A. (2021). Tools, techniques and underground networks of Yahoo-boys in Ibadan City, Nigeria. International Journal of Criminal Justice 3, 99–122. https://doi.org/10.36889/IJCJ.2021.003

120 Chukwuma, O.K. (2024). Understanding the crime-grid of the Nigerian Yahoo boys. National Journal of Cyber Security Law 7(2). https://lawjournals.celnet.in/index.php/njcsl/article/view/1651

2) La segunda fase es la inducción al pago (Billing), una vez consolidado el lazo afectivo, se simulan emergencias que justifican el envío de dinero. Las historias suelen implicar tratamientos médicos, trámites aduaneros, rescates legales o bloqueos bancarios. Para sostener estas ficciones se emplean documentos falsos, audios con voces clonadas y videollamadas deepfake. El objetivo es generar una narrativa emocional coherente con la identidad falsa, y maximizar la disposición de la víctima al sacrificio económico. 3) La tercera fase introduce la coerción emocional: si la víctima sospecha o se niega a continuar, se inicia un proceso de chantaje. Este puede incluir la difusión de fotografías íntimas, la amenaza de exposición en medios falsificados o el contacto con familiares utilizando cuentas suplantadas. La amenaza ya no es física, sino simbólica: se amenaza con destruir la imagen, la vida social o la reputación de la víctima.

Finalmente, está la fase de 4) retiro, los fondos obtenidos se canalizan mediante criptomonedas, tarjetas de regalo o mulas digitales. Una vez completado el fraude, la identidad falsa se abandona y el ciclo comienza nuevamente con otra víctima. Esta lógica de rotación constante permite a los Yahoo Boys mantener su anonimato operativo y dificulta la trazabilidad judicial. Además, alimenta un mercado paralelo de recursos delictivos: scripts de conversación, perfiles clonados, kits de extorsión, manuales de manipulación afectiva y software especializado. Todo esto configura una economía criminal peer-to-peer que combina racionalidad tecnológica, explotación emocional y anonimato algorítmico.

Beneficiarios y víctimas

El perfil de las víctimas es tan diverso como las narrativas utilizadas. Mujeres mayores, personas viudas, migrantes solitarios, adolescentes o adultos mayores con baja alfabetización digital, empresarios en situación de estrés o jóvenes en búsqueda de pertenencia emocional[121]. La violencia aquí es invisible, pero profunda: destruye relaciones familiares, consume recursos económicos, desencadena cuadros depresivos y, en algunos casos, induce al suicidio. A diferencia de los ataques físicos, el daño se perpetúa en el tiempo porque afecta la subjetividad misma de la víctima. La sensación de haber sido engañado en el plano del amor, de la confianza o del deseo no se supera con facilidad. Del mismo modo, esta vergüenza o humillación promueve que en la mayoría de las veces las víctimas no denuncien este tipo de delitos.

121 AIID. (2025). Incident 912: Yahoo boys and scammers from Morocco allegedly target U.S. widows and vulnerable individuals with 'Artificial Patriot' scams. https://incidentdatabase.ai/cite/912

En cuanto a los beneficiarios, el colectivo de los Yahoo Boys no se limita a los estafadores visibles. Existe una economía criminal más amplia, que incluye entrenadores en Telegram, proveedores de identidades sintéticas, diseñadores de deepfakes, programadores de bots y especialistas en evasión digital[122]. Estos actores no están organizados en una jerarquía tradicional, sino que operan como nodos funcionales de una red que se expande por demanda. Cada segmento aporta un recurso, una técnica o un servicio, y a cambio recibe una parte del botín. Este modelo distribuido reduce el riesgo individual, incrementa la eficiencia y promueve la reproducción constante del esquema delictivo.

122 AIID. (2025). Incident 913: Yahoo boys allegedly using AI-generated news videos to blackmail sextortion victims. https://incidentdatabase.ai/cite/913

El uso de IA ha transformado a los Yahoo Boys en un caso paradigmático de lo que podría denominarse criminalidad emocional automatizada. A diferencia de cárteles armados o insurgencias doctrinarias, este grupo no busca controlar territorio, sino narrativas. Su poder no reside en la fuerza, sino en la capacidad de manipular la subjetividad a través del engaño algorítmico. Lo que se roba no es solo el dinero, sino la identidad emocional, el tiempo afectivo, la intimidad simbólica de la víctima.

CASO 5. EL SINDICATO DEL PISO 13 DE POIPET (CAMBOYA)

Desde Poipet, una ciudad fronteriza ubicada en el noroeste de Camboya, justo frente a la provincia tailandesa de Sa Kaeo, emergió entre 2024 y 2025 una de las expresiones más perturbadoras del crimen algorítmico del sudeste asiático: el Sindicato del Piso 13. Poipet, conocida por ser una puerta informal entre Camboya y Tailandia, y un nodo de tránsito para migrantes, apuestas ilegales y comercio gris, se ha convertido en un terreno fértil para operaciones criminales de nueva generación.

En este contexto, el Sindicato del Piso 13, como se ha decidido denominar simbólicamente a esta organización, operaba desde un rascacielos de 18 pisos, donde, en su planta número trece, se desarrollaba una plataforma criminal algorítmica. Desde una sala climatizada y supervisada por operadores subordinados a líderes chinos, se emitían órdenes judiciales simuladas, audiencias digitales fabricadas y detenciones virtuales que terminaban en depósitos bancarios reales[123]. Su sofisticación no residía en la violencia explícita, sino en la capacidad de fingir la legalidad con precisión milimétrica.

El Sindicato del Piso 13 fue parcialmente desarticulado luego del arresto de Ramil Pantawong y Thanawut Kanyaphan[124], dos ciudadanos tailandeses que fungían como operadores intermedios dentro de un ecosistema más vasto, compuesto por redes criminales con sede en China, plataformas tecnológicas para fraude digital, proveedores de software de deepfake, y estructuras de trata de personas utilizadas para el reclutamiento y confinamiento de operadores.

Tecnologías utilizadas

En el corazón técnico del Sindicato del Piso 13 se encontraba un sistema ensamblado con herramientas de GenIA, deepfakes y software de suplantación judicial. Los operadores utilizaban clonadores de voz y modelos de video sintético entrenados para reproducir fielmente a fiscales, jueces y policías tailandeses. En las videollamadas, el interlocutor aparecía con uniforme, entonación, títulos oficiales y una narrativa jurídica perfectamente estructurada. La verosimilitud no era un accesorio: era el dispositivo mismo del delito[125].

Los guiones, redactados originalmente en chino y luego traducidos por intérpretes locales al tailandés, eran cargados en interfaces automatizadas que reproducían una escenografía judicial casi indistinguible de la real. Algunas víctimas, como la modelo y reina de belleza Charlotte Austin, de nacionalidad británico-tailandesa, fueron obligadas a transferir más de 4 millones de baht tailandeses (aproximadamente 112,000 dólares), convencidas de que estaban bajo investigación por lavado de dinero[126].

123 AIID. (2025). Incident 918: AI-aided scam in Thailand allegedly impersonates police to defraud 163 victims. https://incidentdatabase.ai/cite/918

124 Bangkok Post. (2025, March 2). Two men arrested for alleged B4m AI-aided scam against beauty queen. https://www.bangkokpost.com

125 Narim, K. (2025, February 24). Cambodian police raid scam centers in Poipet, discover over 200 foreigners. CamboJA News. https://cambojanews.com

126 THAI.NEWS. (2025, February 3). Charlotte Austin's 4 million baht loss: Inside the Poipet call scam bust in 2025. https://thai.news/news/thailand/charlotte-austins-4-million-baht-loss-inside-the-poipet-call-scam-bust-in-2025

Pero la tecnología no se limitaba al engaño visual. La red desplegaba además apps falsas, portales clonados de organismos estatales y plataformas de pago disfrazadas de cuentas gubernamentales, todo ello alojado en servidores espejo protegidos por VPNs personalizados y redes anónimas como Tor y ZeroNet[127]. Se trataba de una simulación completa del aparato estatal, con todas las capas de su autoridad: desde el timbre de voz hasta el QR de autenticidad.

Modus operandi

El mecanismo fraudulento del Sindicato del Piso 13 puede entenderse como una dramaturgia del poder digital. Todo comenzaba con una llamada o mensaje personalizado: el tono era urgente, la voz impecable. A la víctima se le informaba que estaba bajo sospecha penal. A partir de ahí, se activaba un protocolo perfectamente coreografiado: una videollamada con un "juez", el envío de documentos "oficiales" vía correo electrónico, la amenaza de una orden de aprehensión. La víctima no podía distinguir entre una simulación algorítmica y una autoridad real[128].

Una vez capturada emocionalmente, se le ofrecía una alternativa: transferir sus ahorros a una "cuenta de garantía del Estado" para demostrar buena fe. Las víctimas eran guiadas, paso a paso, por operadores que no levantaban sospechas: hablaban con cortesía, usaban lenguaje técnico, daban tiempos límite como si se tratara de un expediente real[129].

La violencia, en este caso, era de orden simbólico. Ninguna pistola apuntaba a la víctima. Pero sí un sistema algorítmico que replicaba toda la estructura del castigo: la voz del fiscal, el documento sellado, la aplicación que mostraba su nombre como "investigado". La amenaza no era física, era institucional. Pero su eficacia era total.

Las operaciones transfronterizas llevadas a cabo entre Tailandia y Camboya en febrero y marzo de 2025 permitieron el rescate de más de 200 personas, muchas de ellas bajo condiciones de servidumbre digital[130]. No obstante, las detenciones fueron episódicas: la red sigue viva en sus protocolos, sus modelos y sus códigos. El daño no fue sólo financiero. Fue epistémico. Se robó algo más profundo: la capacidad de reconocer al Estado real frente al Estado simulado.

Víctimas y beneficiarios

Las víctimas directas fueron más de 160 personas en Tailandia, con especial incidencia en mujeres jóvenes, empresarios con doble nacionalidad y figuras mediáticas[131]. Sin embargo, la línea entre víctima y victimario fue deliberadamente difuminada. En el mismo piso 13 del edificio en Poipet, donde se grababan las videollamadas falsas, operaban decenas de jóvenes tailandeses que habían sido reclutados con promesas de empleo remoto. Una vez en el lugar, les retiraban el pasaporte, les asignaban tareas específicas, y eran vigilados bajo esquemas de coerción digital[132].

El ecosistema económico de Sindicato del Piso 13 incluía desarrolladores de modelos IA adaptados para suplantación (dark LLMs), proveedores de servidores clandestinos, moderadores de plataformas, diseñadores de interfaces falsas y "mulás" financieras que canalizaban los ingresos por criptomonedas. La plataforma no era simplemente un centro de estafas: era una economía criminal donde se comercializaba la obediencia, se tercerizaba la extorsión y se profesionalizaba el engaño.

Como señala el Penang Institute, este modelo de "fábrica de fraude" se ha expandido por regiones de soberanía débil, como Poipet, Sihanoukville y KK Park, donde la convergencia entre crimen organizado, tecnologías disruptivas y trata de personas ha dado lugar a enclaves de gobernanza criminal autónoma[133].

127 Penang Institute. (2023). Combating scam syndicates in Malaysia and Southeast Asia. Penang Institute Policy Brief. https://penanginstitute.org/publications/policy/combating-scam-syndicates-in-malaysia-and-southeast-asia

128 Raksmey, H. (2025, February 24). Poipet scam compound raids net 230 foreigners, more rescued. The Phnom Penh Post. https://www.phnompenhpost.com/national/poipet-scam-compound-raids-net-230-foreigners-more-rescued

129 Chheng, N. (2025, March 25). National police capture Thai ringleaders during Poipet scam raids. The Phnom Penh Post. https://www.phnompenhpost.com/national/national-police-capture-thai-ringleaders-during-poipet-scam-raids

130 Kiripost. (2025, March 26). Raids on Poipet scam centres find 63 Thais involved in online fraud. https://kiripost.com/stories/cambodia-raids-on-poipet-scam-centres-thais-involved-onine-fraud

131 AIID. (2025). Incident 918: AI-aided scam in Thailand allegedly impersonates police to defraud 163 victims. https://incidentdatabase.ai/cite/918

132 The Nation Thailand. (2025, March 3). 119 Thais from Poipet: Victims or accomplices in a call centre scam? https://www.nationthailand.com/news/policy/40046983

133 Penang Institute. (2023). Combating scam syndicates in Malaysia and Southeast Asia. https://penanginstitute.org/publications/policy

CASO 6. OPERACIÓN CUMBERLAND

La Operación Cumberland representa el primer operativo multinacional centrado en redes criminales que distribuyen y monetizan contenido de abuso sexual infantil generado por AI-CSAM. El caso emerge como un parteaguas histórico, al delinear por primera vez un ecosistema criminal donde la víctima desaparece físicamente, pero sobrevive sintéticamente, replicada por GenIA sin salvaguardas[134].

La operación fue detonada por la detención de un ciudadano danés en noviembre de 2024, quien dirigía una plataforma digital cerrada donde se ofrecían imágenes y videos hiperrealistas de abuso sexual infantil[135]. Este individuo funcionaba como un broker algorítmico: no solo administraba la plataforma, sino que facilitaba la producción, el acceso por suscripción y la circulación mediante criptomonedas. Más de 273 sospechosos fueron identificados en 19 países, lo que refleja la escala transnacional del fenómeno. El operativo implicó a Europol, Interpol y autoridades nacionales en Europa, Oceanía y América, lo que marca un precedente en el tratamiento judicial de delitos cometidos a través de IA[136].

Tecnologías utilizadas

El corazón tecnológico de Cumberland radicó en el uso sistemático de modelos generativos de imágenes por IA entrenados, en algunos casos, con datasets de abuso real. Estas herramientas —entre ellas modelos de código abierto como DeepSeek V3 y variantes manipuladas de LLMs comerciales— eran capaces de generar representaciones visuales de niños en situaciones sexualizadas, sin que mediara contacto físico con una víctima[137].

El perfeccionamiento de los generadores oscuros (dark LLMs) permitió que los ofensores operaran con un grado de especificidad sin precedentes. Las imágenes generadas simulaban entornos altamente reconocibles: habitaciones escolares, parques infantiles y espacios familiares, lo que multiplicaba su impacto simbólico. Al prescindir de la fotografía tradicional, los agresores eliminaron las barreras legales y técnicas para producir CSAM. La IA se convirtió así en la nueva herramienta de industrialización del abuso.

Además, se descubrió que algunos de estos modelos eran diseñados deliberadamente para evadir filtros de seguridad. La manipulación del código fuente y la eliminación de restricciones de contenido explícito permitía a los usuarios acceder a una zona libre de inhibiciones algorítmicas. Esta configuración se complementaba con plataformas de acceso cerrado, autenticación multifactor, y uso de redes anónimas como Tor o ZeroNet, facilitando una circulación impermeable al rastreo convencional[138]. El volumen de archivos generados en apenas meses sugiere que la IA no solo facilitó el delito: lo escaló a niveles industriales.

Modus operandi

El funcionamiento de la red desmantelada bajo la Operación Cumberland puede describirse como una combinación de automatización delictiva y economía criminal simbólica. Todo comenzaba con la creación algorítmica del contenido: los usuarios describían lo que deseaban ver y los dark LLMs producían imágenes sintéticas en serie. Estas imágenes eran luego clasificadas, etiquetadas y almacenadas para ser distribuidas.

La plataforma —cuyo nombre no fue revelado— funcionaba mediante un modelo de suscripción escalonado: los usuarios pagaban tarifas simbólicas o usaban tarjetas prepagadas para acceder a niveles diferenciados de contenido[139]. En paralelo, se documentó el uso de soluciones tecnológicas “pay-as-you-go” que permitían el live-streaming de abuso infantil bajo demanda, una práctica en auge por su rentabilidad, facilidad técnica y bajo riesgo percibido. Además, la infraestructura técnica se apoyaba en redes peer-to-peer y la darknet, que siguen siendo los principales entornos para la circulación no comercial de CSAM, proporcionando anonimato, persistencia de contenidos y un sentido de comunidad entre los ofensores[140]. Esta combinación de medios configuró un ecosistema de abuso algorítmico y de baja fricción, que maximiza tanto la escalabilidad como la impunidad.

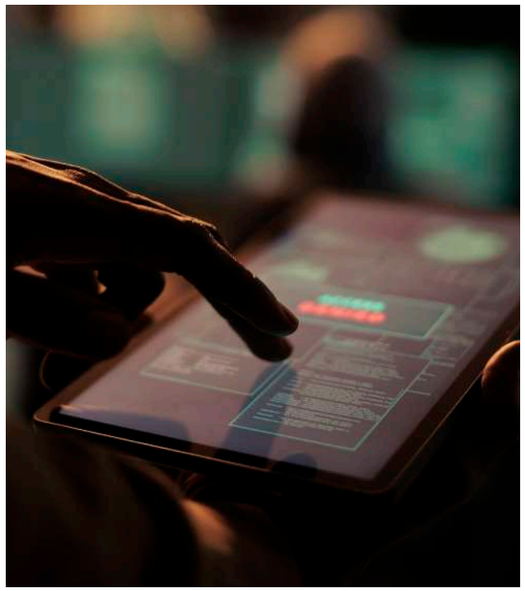
Se documentó además una capa de sofisticación financiera. El uso de criptomonedas como Monero y plataformas de transacción deslocalizadas hacía casi imposible rastrear los flujos monetarios. Al tratarse de contenido generado por IA, los acusados se escudaban en vacíos legales, dificultando su procesamiento judicial. Cumberland, por tanto, no fue solo una red de distribución de imágenes ilegales: fue una zona gris entre lo penalmente sancionable y lo tecnológicamente permisible.

Beneficiarios y víctimas

Los beneficiarios de este ecosistema criminal no eran solo los consumidores directos. Involucraban también a desarrolladores de dark LLMs, programadores que adaptaban modelos para usos ilícitos, moderadores de foros y brokers que vendían accesos y herramientas. Este modelo da forma a una modalidad emergente de criminalidad digital: CSAM-as-a-Service, una plataforma algorítmica descentralizada y monetizable[141].

Las víctimas, por otro lado, no pueden definirse únicamente por la presencia física en el contenido. Existen al menos tres niveles. Primero, los niños cuyas imágenes reales fueron utilizadas para entrenar modelos sin consentimiento. Segundo, los menores cosificados simbólicamente por las imágenes generadas, cuya infancia es convertida en fetiche digital. Tercero, la sociedad en su conjunto, que enfrenta una erosión de sus categorías éticas, legales y afectivas ante la producción masiva de “delitos sin víctimas identificables”.

A largo plazo, el consumo repetido de estas imágenes podría banalizar la violencia sexual contra menores y normalizar una cultura de abuso algorítmico. Este fenómeno reconfigura el daño simbólico: ya no depende del cuerpo violado, sino del imaginario compartido. La Operación Cumberland redefinió la frontera entre lo legal y lo intolerable. Por primera vez, múltiples jurisdicciones reconocen que el CSAM sintético puede y debe ser penalizado, incluso en ausencia de una víctima física. El caso subraya el vacío normativo global, y ha motivado a la Comisión Europea a impulsar una directiva que armonice la legislación en torno a este fenómeno.



IMPLICACIONES ESTRATÉGICAS

Las redes distribuidas de criminalidad algorítmica representan una mutación estructural en el ecosistema delictivo global. A diferencia de las organizaciones jerárquicas tradicionales, cuya lógica se sostiene en el control territorial y la verticalidad del mando, los actores analizados en este bloque operan mediante una arquitectura reticular, transitoria y profundamente adaptativa. Lo que emerge aquí no es una red con centro, sino una serie de nodos interconectados que se reconfiguran en función de la oportunidad, el recurso disponible o el vacío institucional. Esta plasticidad criminal no es una debilidad, sino su mayor fortaleza. La ausencia de jerarquía permite disolver responsabilidades, fragmentar la trazabilidad y escalar la operación sin necesidad de cohesión permanente.

En el caso de Funk-Sec, por ejemplo, no asistimos a la aparición de un nuevo cartel digital, sino al ensamblaje episódico de identidades, habilidades y herramientas que convergen alrededor de objetivos operativos fluidos: ataques informativos, extorsión automatizada, sabotaje reputacional. Lo mismo ocurre con redes como el Clan San Roque o los nodos funcionales en la región del Mekong. Lo que los articula no es una ideología común, ni siquiera una lógica de mercado estable, sino una práctica: la explotación instrumental de tecnologías de IA para fines delictivos, mediada por relaciones efímeras de intercambio, cooperación y anonimato.

134 Nicholls, C. (2025, February 28). Dozens arrested in crackdown on AI-generated child sexual abuse material. CNN. https://edition.cnn.com/2025/02/28/world/ai-child-sex-abuse-europol-operation-intl

135 AIID. (2025). Incident 958: Europol Operation Cumberland investigates at least 273 suspects in 19 countries for AI-generated child sexual abuse material. https://incidentdatabase.ai/cite/958

136 Europol. (2025). Child sexual exploitation. European Union Agency for Law Enforcement Cooperation. https://www.europol.europa.eu/crime-areas/child-sexual-exploitation

137 Burton, J., Janjeva, A., Moseley, S., Alice. (2025). AI and serious online crime. Centre for Emerging Technology and Security (CETaS), The Alan Turing Institute. https://cetas.turing.ac.uk/publications/ai-and-serious-online-crime

138 AIID. (2025). Incident 958: Europol Operation Cumberland investigates at least 273 suspects in 19 countries for AI-generated child sexual abuse material. https://incidentdatabase.ai/cite/958

139 Europol. (2025). Child sexual exploitation. European Union Agency for Law Enforcement Cooperation. https://www.europol.europa.eu/crime-areas/child-sexual-exploitation

140 Burton, J., Janjeva, A., Moseley, S., Alice. (2025). AI and serious online crime. Centre for Emerging Technology and Security (CETaS), The Alan Turing Institute. https://cetas.turing.ac.uk/publications/ai-and-serious-online-crime

141 Nicholls, C. (2025, February 28). Dozens arrested in crackdown on AI-generated child sexual abuse material. CNN. https://edition.cnn.com/2025/02/28/world/ai-child-sex-abuse-europol-operation-intl

La criminalidad, en estos casos, es un servicio distribuido, no una organización estructurada. Tal como ocurre con los Yahoo Boys nigerianos, que combinan estafas románticas, suplantaciones institucionales y herramientas de IA generativa sin necesidad de jerarquías estables, estos actores crean economías algorítmicas del engaño desde la informalidad.

Este desplazamiento tiene consecuencias estratégicas profundas. En primer lugar, exige abandonar la idea del crimen como fenómeno exclusivamente físico o territorial. Las redes distribuidas funcionan como sistemas de ocupación simbólica: ocupan flujos, espacios digitales, imaginarios sociales. El daño que generan —emocional, financiero, cognitivo— no requiere presencia armada ni control espacial. Un solo nodo puede inducir al suicidio financiero de una víctima en Lima, ejecutar una campaña de desinformación en Tegucigalpa, y simular órdenes de embargo en Ciudad de México, sin que sus operadores se encuentren en el continente.

La dislocación entre acto, víctima y perpetrador convierte a estas redes en entes evanescentes, pero no por ello menos letales. El caso de la Operación Cumberland confirma este punto: miles de imágenes sintéticas de abuso infantil fueron producidas y distribuidas por actores que jamás vieron a una víctima física, pero generaron daño masivo desde un enjambre algorítmico sin rostro.

En segundo lugar, el uso de IA como infraestructura operativa de estas redes plantea un desafío epistemológico. La IA ya no es un medio, sino un actor. Interactúa, adapta, persuade, engaña. En Poipet, por ejemplo, no hay un líder que ordena, sino un sistema que automatiza la coerción: dashboards que monitorean productividad, bots que inducen al engaño afectivo, algoritmos que deciden quién vive y quién es transferido a centros más violentos. Esta automatización del poder no solo reduce la exposición de los jefes reales, sino que erosiona las bases mismas de la rendición de cuentas.

¿Quién es responsable cuando el perpetrador es un sistema? ¿Cómo se imputa responsabilidad cuando el daño es generado por una secuencia algorítmica ejecutada en tiempo real por un trabajador esclavizado? Lo mismo ocurre con las redes de montadeudas en México, donde los bots de cobranza, las amenazas deepfake y la manipulación de contactos actúan como un sistema algorítmico de extorsión constante, sin que se pueda identificar a un solo agresor físico.

En tercer lugar, la fragmentación de estas redes erosiona las posibilidades de atribución legal y de cooperación internacional. Las fiscalías y policías están diseñadas para perseguir organizaciones, no ecosistemas. Cada vez que un país desmantela un nodo, surgen tres más en otros contextos, con otros lenguajes, otras tecnologías, otros fines. Esta resiliencia es estructural. No es posible "decapitar" redes distribuidas, porque no tienen cabeza. Lo que se requiere es una transformación estratégica del enfoque: de la persecución de actores aislados al desmantelamiento de infraestructuras técnicas, al mapeo de patrones operativos, a la comprensión de los ciclos de vida de estos sistemas.

A nivel institucional, esta mutación redefine también las nociones de protección y prevención. No basta con fortalecer el perímetro digital; hay que intervenir los contextos de vulnerabilidad que alimentan estas redes. Funk-Sec recluta no solo por habilidad, sino por frustración, marginación o desconfianza sistémica. Poipet opera no solo por brutalidad, sino por precariedad globalizada y complicidad estatal. El Clan San Roque emerge en vacíos de gobernanza donde la tecnología llega antes que el Estado.

En todos los casos, la tecnología criminal es un síntoma, no una causa: es la herramienta que activa condiciones preexistentes de exclusión, impunidad y desesperanza. Las redes distribuidas no son el futuro del crimen. Son su presente. Y si no aprendemos a leer sus patrones, no serán solo más difíciles de perseguir: serán imposibles de entender.

PLATAFORMAS CRIMINALES AUTÓNOMAS (CRIME-AS-A-SERVICE)

En el tránsito del crimen organizado tradicional hacia estructuras digitalizadas, hemos asistido a múltiples transformaciones en la forma de operar, reclutar, comunicar y atacar. Algunas organizaciones jerárquicas adaptaron la IA como extensión del mando. Otras, distribuidas y horizontales, la integraron como recurso para expandir la disrupción y el caos. Pero hay una tercera vía, más reciente y aún más preocupante: las plataformas criminales autónomas, sistemas algorítmicos de alta sofisticación que no operan como colectivos criminales convencionales, sino como infraestructuras funcionales, replicables y estratégicamente diseñadas para facilitar actividades ilícitas a escala global.

Estos modelos no obedecen a jerarquías visibles, ni dependen de liderazgos ideológicos o dinámicas de enjambre. Operan como sistemas encapsulados de criminalidad algorítmica, listos para ser ejecutados por cualquier actor con acceso, sin importar su afiliación o conocimiento técnico. Su interfaz imita la de un asistente de IA comercial, pero su lógica de fondo responde a principios radicalmente distintos: no filtrar, no prevenir, no censurar. Solo ejecutar.

Los tres casos que componen este bloque — Dark LLMs (WormGPT, FraudGPT, DarkBARD), Storm-2139 y Xanthorox AI— representan la materialización de esta nueva lógica criminal. No se trata de organizaciones con miembros, ni de cibercolectivos con causas compartidas. Son modelos que emulan agencia, pero que han sido diseñados para permitir y facilitar prácticas delictivas desde el momento mismo de su concepción. Sus usuarios no requieren saber programar, ni tener acceso a foros sofisticados. Solo necesitan escribir una instrucción

clara —por ejemplo, "crea un correo de extorsión médica"— y el sistema se encargará del resto.

El motivo por el que se agrupan en este bloque no es meramente funcional. No solo comparten capacidades técnicas o patrones de uso. Lo que los une es algo más profundo: una ontología criminal basada en el lenguaje y automatizada por diseño. A diferencia de los deepfakes visuales o los malwares tradicionales, los Dark LLMs operan a través del texto como vector de ataque. El lenguaje no es solo un medio; es el arma. Lo que antes era una herramienta de comunicación, en este contexto se convierte en infraestructura de daño.

En los casos anteriores —ISIS, CJNG, KK Park, Yahoo Boys, FunkSec, el Clan San Roque o las bandas de Montadeudas— la IA funcionaba como complemento o amplificador. En los casos que nos ocupan ahora, la inteligencia artificial es el actor. O más precisamente: es la arquitectura algorítmica sobre la que se montan múltiples actores para escalar sus capacidades ofensivas. No tienen cuerpo, no tienen voz, no tienen ideología. Pero tienen sintaxis, memoria, entrenamiento, y una voluntad programada: asistir en la comisión de delitos. Esta despersonalización del crimen —su reducción a interfaz y su expansión a escala— marca un quiebre cualitativo con todo lo anterior.

Una de las razones fundamentales que justifica el análisis conjunto de los tres casos es que comparten cuatro características. 1) En primer lugar, destacan por su autonomía operativa y su arquitectura modular, lo que significa que no requieren intervención técnica por parte del usuario para ejecutar acciones delictivas. A diferencia de los modelos más simples que solo generan contenido textual (como correos de phishing o malware básico), estas plataformas convierten una simple intención —como extorsionar o sabotear— en una secuencia operativa autónoma.

2) En segundo lugar, ambas plataformas operan bajo una lógica de neutralidad algorítmica y adaptabilidad contextual. Es decir, no tienen una finalidad delictiva única ni filtran sus respuestas por criterios éticos o legales. Ejecutan lo que se les solicita, adaptándose con facilidad a distintos contextos geográficos, jurídicos o lingüísticos.

3) Un tercer rasgo compartido es su diseño orientado al anonimato y la evasión de trazabilidad. Aunque los Dark LLMs ya operan en espacios cifrados con pseudonimato, estas nuevas plataformas llevan la opacidad a un nivel estructural. Incorporan servidores onion, rotación de direcciones IP, spoofing y dominios radicados en jurisdicciones laxas para dificultar su localización y atribución.

4) Finalmente, los evidencian una escalabilidad industrial y una lógica empresarial que trasciende la economía criminal tradicional. Inspiradas en el modelo *Crime-as-a-Service* (*Caas*), estas herramientas se presentan como productos tecnológicos legítimos, con documentación técnica, niveles de suscripción, soporte personalizado y comunidades de usuarios.

En ese sentido, el estudio de estos modelos desborda la ciberseguridad. Exige pensar en nuevas categorías para el análisis criminológico, nuevos marcos regulatorios, y nuevas respuestas institucionales. Si el crimen se convierte en servicio algorítmico, y si el lenguaje se convierte en arma automatizada, entonces el campo de batalla ya no es solo la red oscura, sino el texto mismo: el prompt, la conversación, el discurso generado.

CASO 1. DARK LLMS (WORMGPT, FRAUDGPT, DARKBARD)

Los denominados Dark LLMs representan una inflexión en la arquitectura del crimen digital contemporáneo. Lejos de estar estructurados como redes jerárquicas o células distribuidas, se constituyen como infraestructuras algorítmicas autónomas, replicables y altamente adaptativas. Su despliegue no requiere territorio, jerarquía ni coordinación humana compleja: basta con un modelo entrenado fuera de los márgenes normativos, una interfaz amigable, y un mercado clandestino dispuesto a pagar por sus servicios.

Su modularidad permite que actores sin conocimientos técnicos realicen tareas ofensivas complejas, como generar malware, diseñar campañas de phishing o crear deepfakes personalizados, todo mediante lenguaje natural. A la fecha, se han detectado más de 212 variantes de modelos maliciosos activos en plataformas clandestinas, incluyendo WormGPT, XXXGPT, WolfGPT y GhostGPT, lo que confirma la rápida expansión de este fenómeno[142].

Modelos como WormGPT, FraudGPT y DarkBARD han sido detectados en plataformas cifradas de la *darknet*, en canales como BreachForums y grupos privados de Telegram, donde son promocionados como herramientas de ciberataque accesibles y sin restricciones[143]. Ofrecen paquetes escalables, suscripciones premium y soporte técnico en tiempo real, lo que refuerza su lógica de plataforma y consolida su rol como servicios criminales *"as-a-Service"*. Su aparición no fue espontánea: deriva de un ecosistema de entrenamiento paralelo que se alimenta de modelos filtrados como GPT-J, LLaMA o Codex, los cuales son reconfigurados para eliminar cualquier tipo de censura algorítmica[144].

A diferencia de las herramientas comerciales, estos LLMs son entrenados con datasets extraídos ilícitamente mediante técnicas de *web scraping* masivo o filtraciones en repositorios públicos como GitHub, violando así tanto los derechos de autor como principios éticos de la IA[145]. El objetivo no es solo técnico, sino económico: construir asistentes delictivos listos para operar en mercados sin regulación. Además, utilizan técnicas de jailbreaking como *prompt injection* o *role-play inversion* para desactivar salvaguardas éticas, facilitando la producción de contenidos prohibidos sin supervisión[146]. El objetivo no es solo técnico, sino económico: construir asistentes delictivos listos para operar en mercados sin regulación, que democratizan el acceso a capacidades ofensivas avanzadas.

Tecnologías utilizadas

La arquitectura de estos modelos se basa en una lógica de maximización ofensiva algorítmica. WormGPT, FraudGPT y DarkBARD, por ejemplo, permiten la creación de campañas de *spear phishing* en lenguaje natural, la generación automatizada de malware evasivo, la falsificación de documentos y la manipulación de interfaces gráficas sin necesidad de intervención humana directa[147]. Estos modelos han sido descritos como "sin filtros, sin límites morales", debido a su capacidad para generar contenidos que incluyen amenazas, discursos de odio, manuales de hackeo, instrucciones para delitos financieros y más. En particular, su carácter "*uncensored*" —sin mecanismos de moderación activa— es clave para su explotación criminal, lo que ha llevado a que múltiples investigadores los clasifiquen como herramientas de alto riesgo para la seguridad digital global[148].

Un elemento técnico clave en su funcionamiento es el uso sistemático de *jailbreaking*. Entre las estrategias más frecuentes se encuentran el *prompt injection*, el uso de *tokens* de escape, *encoding inverso* y *chain-of-thought manipulation*, técnicas todas ellas diseñadas para sortear las barreras de moderación algorítmica embebidas en los modelos base[149]. Originalmente concebidas como herramientas de auditoría ética, estas técnicas han sido adaptadas y automatizadas para su uso delictivo.

Además, los desarrolladores de *Dark LLMs* han incorporado funciones avanzadas como verificación de calidad algorítmica, que permite ajustar las respuestas generadas según criterios

142 CybelAngel. (2025). Gen AI and the rise of uncensored LLMs on the dark web. CybelAngel. https://cybelangel.com/gen-ai-uncensored-llms

143 Barman, D., Guo, Z., Conlan, O. (2024). The dark side of language models: Exploring the potential of LLMs in multimedia disinformation generation and dissemination. Machine Learning with Applications. https://doi.org/10.1016/j.mlwa.2024.100545

144 Schultz, J. (2024, junio 4). Cybercriminal abuse of large language models. Talos Intelligence. Cisco Talos. https://blog.talosintelligence.com/cybercriminal-abuse-of-large-language-models/

145 Anggorojati, B., Perdana, A., Wijaya, D. (2024, July 24). FraudGPT and other malicious AIs are the new frontier of online threats. What can we do? The Conversation. https://theconversation.com/fraudgpt-and-other-malicious-ais-are-the-new-frontier-of-online-threats-what-can-we-do-234820

146 Vongthongsri, K. (2025, March 15). How to jailbreak LLMs one step at a time: Top techniques and strategies. Confident AI. https://www.confident-ai.com/blog/how-to-jailbreak-llms-one-step-at-a-time

147 Ruvnet. (2024). The emergence of malicious large language models (LLMs) and the next frontier of symbolic-AI integration. GitHub. https://gist.github.com/ruvnet/6bd83dcc7dd6e98e86d600ed13576baf

148 Iyer, P. (2024, January 18). Studying underground market for large language models, researchers find OpenAI models power malicious services. Tech Policy Press. https://www.techpolicy.press/studying-black-market-for-large-language-models-researchers-find-openai-models-power-malicious-services/

149 Vongthongsri, K. (2025, March 15). How to jailbreak LLMs one step at a time: Top techniques and strategies. Confident AI. https://www.confident-ai.com/blog/how-to-jailbreak-llms-one-step-at-a-time

específicos definidos por el usuario. Esta adaptación dinámica convierte a estos modelos en verdaderos generadores de contenido a medida, capaces de clonar identidades digitales, simular conversaciones humanas con realismo contextual, o generar scripts que replican interfaces institucionales completas.

Finalmente, investigaciones recientes han documentado que estos modelos no sólo funcionan sobre arquitecturas propias, sino que muchos utilizan como *backend* versiones filtradas o accedidas ilícitamente de modelos comerciales, como GPT-3.5, GPT-4 o Claude-2, vulnerando así las salvaguardas de seguridad originalmente implementadas por sus desarrolladores[150]. Esto demuestra no sólo la sofisticación de las técnicas de acceso no autorizado (*LLMjacking*), sino también la fragilidad estructural del ecosistema algorítmico comercial frente a su manipulación adversaria.

Modus operandi

El modelo operativo de los *Dark LLMs* se inscribe plenamente en la lógica del *Crime-as-a-Service*. Las plataformas se distribuyen bajo un esquema de suscripciones en criptomonedas, con precios que oscilan entre $30 y $200 USD mensuales, dependiendo del nivel de acceso, personalización y soporte técnico[151]. Los usuarios reciben credenciales, acceso a dashboards privados y, en algunos casos, APIs que permiten integrar el modelo en flujos de trabajo delictivos automatizados.

El *onboarding* para nuevos usuarios de Dark LLMs es minimalista y accesible: no requiere conocimientos técnicos, únicamente motivación delictiva. La interfaz se presenta como un asistente conversacional, donde basta con ingresar una instrucción —como redactar un mensaje coercitivo o un guion de estafa— para que el sistema genere el contenido completo, estructurado y adaptado al contexto socio-cultural deseado.

Los desarrolladores han incorporado funciones de ofuscación de *tokens*, rotación de direcciones IP y *scraping* automatizado de datos públicos, lo que permite que los ataques desplegados desde estas plataformas eludan los mecanismos tradicionales de

150 Iyer, P. (2024, January 18). Studying underground market for large language models, researchers find OpenAI models power malicious services. Tech Policy Press. https://www.techpolicy.press/studying-black-market-for-large-language-models-researchers-find-openai-models-power-malicious-services/

151 Poireault, K. (2023). The dark side of generative AI: Five malicious LLMs found on the dark web. Infosecurity Europe. https://www.infosecurityeurope.com/en-gb/blog/threat-vectors/generative-ai-dark-web-bots.html

EL PACCTO 2.0

rastreo y monitoreo[152]. En los casos más avanzados, como los explorados por investigadores en GitHub, se ha documentado el diseño experimental de capacidades neuro-simbólicas, lo que sugiere que algunos modelos podrían razonar, adaptar y priorizar acciones con base en reglas lógicas internas —un paso hacia la integración de IA generativa con estructuras de toma de decisión más complejas, aunque su implementación práctica aún se encuentra en evaluación[153].

Víctimas y beneficiarios

El caso de los Dark LLMs revela un desafío profundo y multivectorial. No se trata de herramientas aisladas, sino de infraestructuras criminales replicables que operan por fuera del marco legal, ético y técnico de la inteligencia artificial. La amenaza no reside únicamente en el contenido que generan, sino en la capacidad de sistematizar el crimen, escalarlo y distribuirlo globalmente con rapidez y bajo costo.

Las víctimas de los Dark LLMs son tan diversas como sus funcionalidades. Desde usuarios individuales engañados por campañas de ingeniería social, hasta empresas atacadas con scripts automatizados, o plataformas institucionales vulneradas mediante deepfakes textuales. También, se ha documentado su uso en la creación de contenido ilegal, extorsión reputacional y campañas de desinformación dirigidas.

En contraste, los beneficiarios incluyen una amplia gama de actores: desde operadores solitarios que monetizan estafas rápidas, hasta brokers financieros, centros de fraude digital, milicias digitales no estatales y colectivos ideológicos interesados en desestabilización política. Estas herramientas han eliminado la barrera técnica del cibercrimen, convirtiendo la intención en acción mediante una interfaz algorítmica accesible.

Las soluciones actuales —como listas negras, filtros de contenido o regulaciones nacionales— resultan insuficientes frente a una amenaza que se reinventa con cada filtración, cada jailbreak exitoso y cada nuevo repositorio explotado. Las respuestas, según los análisis más recientes, deben incluir el desarrollo de sistemas híbridos de detección, auditorías algorítmicas en tiempo real, tipificación penal del *LLMjacking* y colaboración internacional entre desarrolladores, plataformas y marcos jurídicos emergentes.

152 CybelAngel. (2023). The dark side of Gen AI: Uncensored large language models [white paper]. https://cybelangel.com/gen-ai-uncensored-llms

153 Ruvnet. (2024). The emergence of malicious large language models (LLMs) and the next frontier of symbolic-AI integration. GitHub. https://gist.github.com/ruvnet/6bd83dcc7dd6e98e86d600ed13576baf

CASO 2. XANTHOROX AI

A diferencia de los cárteles tradicionales, cuya estructura jerárquica y territorial define sus operaciones, o de las redes de fraude humano-digital híbridas como los Yahoo Boys o los Montadeudas que combinan extorsión telefónica con técnicas digitales, Xanthorox AI no constituye una organización criminal en el sentido clásico. No posee líderes visibles, no está anclada a una geografía, ni se articula en células o franquicias. En cambio, se presenta como una plataforma algorítmica autónoma, cuyo diseño modular, autohospedado y escalable permite su uso por individuos o colectivos sin mediación institucional. Su naturaleza es la de un intermediario automatizado entre el usuario y la ejecución del delito, lo que la sitúa en una nueva categoría de amenaza: no tanto una red, sino una infraestructura replicable de crimen digital[154].

Su irrupción pública ocurrió en el primer trimestre de 2025, cuando fue detectada por comunidades especializadas en inteligencia de amenazas dentro de canales cifrados y foros de la *darknet*[155]. Allí fue anunciada no como un simple bot o software de ataque, sino como una suite de IA alojada en servidores *onion,* con capacidades de autoentrenamiento y adaptación contextual. Esto significa que Xanthorox AI no requiere comandos codificados: puede interpretar lenguaje natural, adaptar sus ataques al perfil de la víctima y emular entornos operativos sin que un operador humano intervenga directamente[156]. Se trata, en esencia, de un sistema algorítmico que aprende, planifica y ejecuta con base en objetivos preestablecidos por el usuario.

Lo que resulta especialmente perturbador es el contexto en el que esta plataforma emerge. Investigaciones de medios como *Scientific American*[157], *The 420*[158] y *SlashNext*[159] coinciden en advertir que Xanthorox tiene el potencial de democratizar el crimen cibernético, permitiendo que actores no estatales, proxies armados y redes financieras ilícitas accedan a capacidades ofensivas comparables a las de una unidad cibernética estatal.

Tecnologías utilizadas

El corazón operativo de Xanthorox AI radica en su arquitectura modular. A diferencia de herramientas como WormGPT o FraudGPT, que se construyen sobre versiones alteradas de modelos comerciales, Xanthorox ha sido desarrollada desde cero utilizando modelos propios y, según se ha filtrado, algunas adaptaciones de LLMs como LLaMA o Codex, obtenidas tras filtraciones en foros de ingeniería inversa[160]. La diferencia no es meramente técnica, sino ontológica: Xanthorox no se limita a "romper" las restricciones éticas de un modelo comercial, sino que incorpora su propia lógica de desarrollo delictivo, alineada con fines ofensivos y operativos.

Uno de los avances más peligrosos de esta plataforma es el uso de interfaces conversacionales en lenguaje natural. Esto implica que un usuario sin conocimientos técnicos puede solicitar, por ejemplo, un ataque de inyección SQL, una campaña de phishing dirigida a un hospital específico, o una manipulación de credenciales biométricas, y la IA genera el código, estructura el ataque y despliega la operación de forma autónoma. El sistema se adapta al tipo de blanco y al nivel de sofisticación requerido, lo que convierte a Xanthorox en un verdadero asistente delictivo por voz y texto.

Entre las funcionalidades más innovadoras se encuentra el desarrollo de exploits a partir de descripciones simples. En teoría, bastaría una orden como "buscar vulnerabilidades en infraestructura hospitalaria de país X" para que el sistema analice puertos abiertos, identifique librerías desactualizadas y genere un script de explotación viable.

De igual forma, mediante técnicas de *morphing*, puede clonar interfaces bancarias completas, emulando comportamiento de usuarios y patrones gráficos con precisión casi forense. Otra función destacada es su módulo de fraude telefónico, que permite configurar voces sintéticas por idioma, acento y género, lo que potencia campañas de suplantación a gran escala, con efectos devastadores en contextos de baja alfabetización digital[161].

154 AIID (2025, April 7). Incident 1015: Reported darknet launch of Xanthorox AI introduces autonomous cyberattack platform. https://incidentdatabase.ai/cite/1015/

155 Griffin, M. (2025, April 26). Revolutionary autonomous cyberattack platform emerges on the dark web. Fanatical Futurist. https://www.fanaticalfuturist.com/2025/04/revolutionary-autonomous-cyberattack-platform-emerges-on-the-dark-web/

156 Ahmed, D. (2025, April 7). Xanthorox AI Surfaces on Dark Web as Full Spectrum Hacking Assistant. Hackread. https://hackread.com/xanthorox-ai-dark-web-full-spectrum-hacking-assistant/

157 Béchard, D. E. (2025, May 7). *Xanthorox AI lets anyone become a cybercriminal*. Scientific American. https://www.scientificamerican.com/article/xanthorox-ai-lets-anyone-become-a-cybercriminal/

158 Nath, S. (2025, April 13). *This AI tool empowers cybercriminals with advanced capabilities—No jailbreaks needed*. The420.in. https://www.the420.in/this-ai-tool-empowers-cybercriminals-with-advanced-capabilities-no-jailbreaks-needed/

159 SlashNext. (2025). Xanthorox AI – The next-gen malicious AI. https://www.slashnext.com/xanthorox-next-gen

160 Ahmed, D. (2025, April 7). Xanthorox AI surfaces on dark web as full spectrum hacking assistant. Hackread. https://hackread.com/xanthorox-ai-dark-web-full-spectrum-hacking-assistant/

161 Béchard, D. E. (2025, May 7). *Xanthorox AI lets anyone become a cybercriminal*. Scientific American. https://www.scientificamerican.com/article/xanthorox-ai-lets-anyone-become-a-cybercriminal/

Modus operandi

El funcionamiento operativo de Xanthorox AI responde a una lógica de sofisticación modular, diseñada para escalar desde usuarios novatos hasta operadores avanzados. Su modelo de negocio se estructura en dos niveles claramente diferenciados. En la capa superficial, la plataforma ofrece acceso gratuito a funcionalidades básicas a través de canales abiertos como Telegram y Discord. En esta modalidad se habilitan herramientas elementales como generadores de correos de phishing, scripts simples de estafa y simulaciones básicas de ingeniería social[162]. Esta estrategia no solo reduce la fricción de entrada para nuevos usuarios, sino que también actúa como un mecanismo de reclutamiento y expansión, ampliando el ecosistema criminal mediante automatización accesible.

Sin embargo, el verdadero poder de Xanthorox se despliega en su versión profesional. Esta ha sido identificada en foros cerrados de la *darknet*, donde se ofrece a operadores con fines claramente ofensivos. Aunque los detalles sobre su acceso y monetización aún no están plenamente documentados, todo indica que su uso se reserva a usuarios con conocimiento de canales específicos dentro del ecosistema clandestino. En este nivel, el usuario accede a una suite ofensiva de espectro completo, construida sobre una arquitectura modular e integrada que permite planear y ejecutar ciberataques sin necesidad de conocimientos técnicos avanzados[163].

La plataforma está compuesta por tres módulos principales, cada uno orientado a una fase específica de la operación ofensiva. 1) El primero, Xanthorox Coder, automatiza la redacción de código malicioso. Su motor permite generar scripts de explotación personalizados, adaptar cargas a vulnerabilidades detectadas en tiempo real y modificar artefactos para evadir firmas tradicionales[164]. Esta capacidad se basa en modelos entrenados para comprender no solo la sintaxis del lenguaje, sino también el contexto operativo del ataque.

2) El segundo módulo, Xanthorox Vision, integra reconocimiento visual avanzado y procesamiento de imágenes mediante redes neuronales convolucionales. Con él, el sistema puede analizar capturas de pantalla, interfaces gráficas, diagramas de redes o formularios digitales, generando réplicas sintéticas capaces de engañar a usuarios reales en ataques de *phishing*, clonación de sitios institucionales o fraude documental[165].

3) El tercero y más inquietante componente, Xanthorox Reasoner Advanced, emula procesos de razonamiento humano y simulación social. A través de este módulo, el sistema interpreta patrones conductuales, propone estrategias adaptativas de ingeniería social y prioriza vectores de ataque con base en el perfil estimado de la víctima[166]. Esta lógica permite que el usuario solo deba formular su intención —por ejemplo, vulnerar la seguridad de una fundación humanitaria— para que la plataforma articule una secuencia completa de acciones: desde la recolección inicial de metadatos hasta la extracción de credenciales, manipulación de documentos y ejecución de campañas de desinformación.

Todo el ciclo operativo está diseñado para evitar trazabilidad. Xanthorox emplea técnicas de *scraping* automatizado, generación de identidades falsas, análisis de puertos y explotación de APIs públicas mal configuradas. La ejecución del ataque puede realizarse desde servidores intermediarios, en redes cifradas o incluso mediante bots programados para activar rutinas tras cierto intervalo temporal. Algunas variantes ya detectadas incluyen funciones que sugieren rutas de evasión legal, como el uso de jurisdicciones sin marcos actualizados de ciberlegislación o el aprovechamiento de tratados bilaterales con vacíos normativos[167].

Beneficiarios y víctimas

El atributo más inquietante de Xanthorox AI no es su capacidad técnica, sino su efecto sociotécnico: eliminar la barrera de entrada al cibercrimen. Lo que antes requería conocimientos avanzados, infraestructura, experiencia y redes de apoyo, hoy se sintetiza en una interfaz donde basta escribir una intención para obtener una operación completa. Esta democratización del delito significa que cualquier individuo con motivación política, financiera o incluso emocional puede lanzar un ataque sofisticado sin depender de un colectivo criminal estructurado.

162 Idem.

163 Kelley, D. (2025, April 7). Xanthorox AI – The next generation of malicious AI threats emerges. SlashNext. https://slashnext.com/blog/xanthorox-ai-the-next-generation-of-malicious-ai-threats-emerges/

164 Griffin, M. (2025, April 26). Revolutionary autonomous cyberattack platform emerges on the dark web. Fanatical Futurist. https://www.fanaticalfuturist.com/2025/04/revolutionary-autonomous-cyberattack-platform-emerges-on-the-dark-web/

165 Ahmed, D. (2025, April 7). Xanthorox AI surfaces on dark web as full spectrum hacking assistant. Hackread. https://hackread.com/xanthorox-ai-dark-web-full-spectrum-hacking-assistant/

166 Kelley, D. (2025, April 7). Xanthorox AI – The next generation of malicious AI threats emerges. SlashNext. https://slashnext.com/blog/xanthorox-ai-the-next-generation-of-malicious-ai-threats-emerges/

167 AIID. (2025, April 7). Incident 1015: Reported darknet launch of Xanthorox AI introduces autonomous cyberattack platform. https://incidentdatabase.ai/cite/1015/

Si bien los reportes técnicos no confirman usuarios específicos, el diseño accesible y modular de Xanthorox AI permite inferir su potencial adopción por actores de diversas regiones, incluidos extorsionadores individuales, milicias digitales no estatales y operadores financieros que buscan evadir controles tradicionales. Esta flexibilidad, unida a su disponibilidad en redes clandestinas, también ha despertado preocupación sobre su eventual uso por parte de actores estatales en escenarios de guerra cognitiva o manipulación electoral.

Las posibles víctimas de Xanthorox AI abarcan un espectro potencialmente amplio, dada la facilidad con la que su diseño permite generar ataques personalizados sin mediación técnica. Si bien no existen reportes públicos que atribuyan su uso a incidentes específicos, las capacidades descritas en su arquitectura —como la automatización de ingeniería social, la falsificación de documentos y el despliegue de scripts maliciosos— podrían afectar desde sistemas financieros locales hasta ONGs, hospitales o plataformas cívicas. Esta amplitud operativa, unida a su disponibilidad en redes clandestinas, refuerza la idea de que Xanthorox no discrimina objetivos: ejecuta aquello que el usuario instruya, lo que plantea un riesgo algorítmico de tipo sistémico.

CASO 3. STORM-2139

Storm-2139 representa un nuevo paradigma en la arquitectura del crimen digital global. No se trata de una red distribuida, ni clásicamente jerárquica, sino de una plataforma criminal autónoma, distribuida, multirregional y altamente tecnificada, que opera en la intersección entre la explotación de vulnerabilidades en sistemas de GenIA y el comercio de acceso ilícito como servicio. Esta forma de operación se ha definido como Crime-as-a-Service 5.0[168].

La red fue identificada por primera vez en 2024 por Microsoft tras una serie de accesos no autorizados a su infraestructura de Azure OpenAI. La investigación judicial posterior permitió reconstruir una estructura tripartita funcional que incluía, por un lado, a los desarrolladores especializados en *jailbreaking* de modelos de lenguaje y en el diseño de herramientas para evadir salvaguardas éticas[169]. En un segundo nivel se encontraban los proveedores o intermediarios, encargados de comercializar accesos robados y herramientas customizadas para una clientela compuesta por actores maliciosos. Finalmente, el tercer nivel lo ocupaban los usuarios, distribuidos internacionalmente, que utilizaban los modelos manipulados para producir deepfakes sexuales, contenido ilegal o narrativas desinformativas[170].

La transición digital de Storm-2139 fue facilitada por el uso de servidores en plataformas descentralizadas como Discord, dominios con extensiones poco reguladas como .to y .ws, y una interfaz semipública que funcionaba como soporte técnico y canal de distribución. Esta infraestructura simulaba una *start-up* tecnológica, lo que permitió operar con eficiencia empresarial sin necesidad de jerarquías formales ni una sede física. La autonomía funcional, combinada con la movilidad digital de sus miembros, convierte a Storm-2139 en un caso ejemplar de cómo los límites entre crimen organizado y ciberplataformas se han vuelto porosos en la era de la IA.

168 Masada, S. (2025, February 27). Disrupting a global cybercrime network abusing generative AI. Microsoft On the Issues. https://blogs.microsoft.com/on-the-issues/2025/02/27/disrupting-cybercrime-abusing-gen-ai/

169 Johnson, D.B. (2025, February 27). Microsoft IDs developers behind alleged generative AI hacking-for-hire scheme. CyberScoop. https://cyberscoop.com/microsoft-generative-ai-azure-hacking-for-hire-amended-complaint/

170 AIID. (2025). Incident 955: Global cybercrime network Storm-2139 allegedly exploits AI to generate deepfake content. https://incidentdatabase.ai/cite/955

Tecnologías utilizadas

El vector técnico central de Storm-2139 fue el llamado *LLMjacking*, es decir, el acceso no autorizado a instancias comerciales de modelos de lenguaje grande (LLM), particularmente aquellos vinculados a OpenAI, mediante el uso de credenciales API robadas[171]. Estas claves de acceso se obtenían a través de repositorios expuestos públicamente, como GitHub, o por negligencia operativa de usuarios que las incluían en códigos no asegurados.

Una vez obtenido el acceso, los operadores de Storm-2139 implementaban herramientas como LLMUnlocker o PromptBypassPro[172]. Estas aplicaciones estaban diseñadas para deshabilitar las funciones de moderación, los filtros de contenido y los controles éticos de los modelos. El resultado era la generación de respuestas sin censura que podían incluir contenido sexual explícito, violencia, incitación al odio y prompts orientados a la fabricación de malware o a la elaboración de discursos manipuladores. Los atacantes integraban también servicios de anonimato y ofuscación de tokens, y ofrecían paquetes con suscripciones premium, soporte técnico y acceso a foros exclusivos para clientes frecuentes.

En sus versiones más sofisticadas, las herramientas desarrolladas por Storm-2139 incluían interfaces visuales para manipular directamente los modelos sin requerir habilidades técnicas avanzadas, democratizando el uso criminal de la IA. Además, se detectaron sistemas de automatización para generar contenido de forma masiva, incluyendo scripts que solicitaban miles de prompts por hora, lo que multiplicaba el daño potencial. Esta infraestructura también incorporaba mecanismos de verificación de calidad algorítmica para asegurar que las respuestas generadas respondieran a las expectativas del mercado ilegal.

Modus operandi

Storm-2139 se consolidó como una sofisticada operación de hacking por encargo, centrada en la explotación algorítmica de sistemas de GenIA. Su cadena operativa comenzaba con el rastreo automatizado de credenciales API expuestas en repositorios públicos o foros clandestinos. Estas

171 Microsoft. (2025, February 29). Microsoft disrupts Storm-2139 for LLMjacking and Azure AI exploitation. https://www.microsoft.com/en-us/security/blog

172 Tharayil, R. (2025, February 28). Microsoft expands legal action against AI abuse network Storm-2139. Tech Monitor. https://www.techmonitor.ai/technology/cybersecurity/microsoft-legal-action-storm-2139

credenciales, una vez verificadas, habilitaban el acceso no autorizado a entornos comerciales de Azure OpenAI, donde los operadores implementaban técnicas de jailbreak para eliminar las barreras normativas impuestas por los desarrolladores[173].

Con los modelos comprometidos, se desplegaban instancias manipuladas que podían ser utilizadas de forma encubierta para generar contenido ilícito. Storm-2139 no solo vendía el acceso a estas versiones adulteradas, sino que ofrecía paquetes escalables con niveles de suscripción, soporte personalizado y herramientas adicionales para integrarlas en flujos de trabajo criminales. Algunas de estas herramientas automatizaban la producción de contenido, permitiendo a los usuarios generar miles de imágenes, narrativas falsas o instrucciones técnicas con una mínima intervención humana.

El mercado clandestino al que Storm-2139 abastecía incluía a clientes interesados en crear deepfakes sexuales, contenidos para chantaje, campañas de desinformación y propaganda electoral. A través de foros cerrados y sistemas de referencia, el grupo fomentaba una lógica de comunidad entre los compradores, promoviendo una cultura de "resistencia cognitiva" frente a lo que describían como censura algorítmica. En esa narrativa, el uso sin restricciones de la IA era presentado como un acto de desobediencia digital y una forma de reapropiación tecnológica frente al dominio de las grandes plataformas[174].

Víctimas y beneficiarios

El impacto de las operaciones de Storm-2139 se extendió a múltiples víctimas. En primer lugar, las personas cuyas imágenes fueron utilizadas para crear contenido sexual explícito sin su consentimiento. Celebridades, periodistas, figuras políticas y activistas fueron objeto de estas agresiones algorítmicas[175]. A ello se suma el daño institucional que sufrió Microsoft, tanto en términos de reputación como de responsabilidad legal y técnica.

Instituciones financieras que procesaron pagos vinculados a los servicios de Storm-2139 también se vieron implicadas, al igual que plataformas como Discord y GitHub que funcionaron como vectores logísticos involuntarios[176]. Del lado de los beneficiarios, se identificaron desde operadores de centros de estafa en el sudeste asiático hasta consultores de imagen digital radicados en Europa del Este, todos ellos interesados en utilizar GenIA sin restricciones para actividades como extorsión, propaganda o manipulación reputacional.

Frente a esto, es urgente impulsar marcos legislativos internacionales que tipifiquen el LLMjacking como delito, y articular redes de cooperación entre desarrolladores, instituciones judiciales y plataformas tecnológicas para la protección de derechos digitales. Microsoft, en sus diversos comunicados, ha subrayado la necesidad de una arquitectura de confianza que no se limite a medidas técnicas, sino que incorpore gobernanza, rendición de cuentas y mecanismos de reparación para las víctimas[177]. Storm-2139 no es un caso aislado, sino el anticipo de una nueva etapa en la evolución del crimen digital algorítmico.

IMPLICACIONES ESTRATÉGICAS

La emergencia de plataformas criminales autónomas como los Dark LLMs, Xanthorox AI y Storm-2139 no constituye una simple sofisticación tecnológica dentro del crimen organizado: representa una mutación ontológica del delito mismo. Ya no hablamos de organizaciones que utilizan tecnología, sino de arquitecturas algorítmicas que operan el crimen sin requerir sujetos visibles, mandos intermedios ni causas ideológicas. Se trata de la consolidación de una nueva ecología delictiva, donde el lenguaje se convierte en vector, el modelo en agente y la interfaz en escenario de daño.

Una de las transformaciones más profundas que se observa en estos casos es la despersonalización del crimen. A diferencia de las organizaciones jerárquicas o las redes distributivas, que requerían intervención humana directa para ejecutar extorsiones, manipular emociones o negociar los términos del chantaje, estas nuevas plataformas eliminan la necesidad del cuerpo, del acento, del acoso presencial. El usuario ya no necesita construir una narrativa: basta con describir una intención. El modelo se encarga de ejecutar la sintaxis de la agresión. Esta eliminación del vínculo directo entre autor y víctima desplaza el campo de responsabilidad penal y desafía los marcos legales que aún dependen de la intención, la autoría y la trazabilidad.

La arquitectura modular de estos sistemas habilita un proceso de escalamiento industrial de la criminalidad, que es radicalmente distinto a lo que se observa en organizaciones tradicionales. *Storm-2139*, por ejemplo, ofrecía suscripciones diferenciadas por nivel de acceso, soporte técnico y volumen de producción de contenido malicioso. Esta lógica de "crimen como servicio" trasciende la informalidad: simula el comportamiento de una startup, con atención al cliente, documentación técnica y modelos de precios. Pero su producto no es una solución tecnológica benigna, sino una interfaz capaz de generar miles de deepfakes sexuales, manuales de sabotaje o campañas de manipulación electoral con una eficiencia que rivaliza con la de cualquier empresa legal de software.

El fenómeno más inquietante es que estas plataformas no necesitan crecer para ser peligrosas. A diferencia de los carteles, que requieren expansión territorial o alianzas para aumentar su influencia, aquí basta con una sola instancia de ejecución para generar daño sistémico. Un modelo filtrado, como los que dieron origen a WormGPT o FraudGPT, puede ser entrenado en la clandestinidad, cargado en servidores onion y desplegado con total anonimato, habilitando a un operador solitario a lanzar campañas de estafa masiva o ingeniería social dirigida sin haber escrito una sola línea de código.

En este ecosistema, la atribución penal se vuelve difusa. No hay líderes visibles, no hay células identificables, no hay reuniones. Solo hay prompts, instrucciones escritas en lenguaje natural que desencadenan secuencias delictivas autónomas. El crimen se convierte en conversación, el delito en *output* algorítmico. Esto desafía no solo las capacidades de ciberinteligencia y forensicidad digital, sino los propios fundamentos de la responsabilidad penal individual. ¿Quién responde cuando el crimen ha sido automatizado y encapsulado en una arquitectura diseñada para evitar cualquier forma de trazabilidad?

Los tres casos presentados revelan también un riesgo emergente que aún no ha sido plenamente comprendido: la posibilidad de hibridación entre actores estatales y plataformas autónomas criminales. Las capacidades ofrecidas por sistemas como Xanthorox AI —clonación de interfaces, suplantación de voz, despliegue de exploits personalizados— son atractivas no solo para estafadores o milicias digitales, sino también para aparatos estatales interesados en operaciones encubiertas, manipulación de opinión pública o represión sin huella. El uso de estos modelos por proxies o contratistas armados de baja trazabilidad incrementa el riesgo de una militarización oculta del crimen algorítmico, en la que la frontera entre guerra cognitiva y crimen cibernético se difumina peligrosamente.

Tampoco puede subestimarse el impacto estructural sobre las víctimas. Las herramientas presentadas no discriminan objetivos. Su diseño orientado a instrucciones abiertas permite que cualquier persona —sin importar edad, género o contexto— pueda convertirse en blanco de extorsión automatizada, suplantación identitaria o manipulación emocional. Mientras las víctimas de los montadeudas aún podían identificar una llamada, un rostro o una cuenta bancaria, las víctimas de Storm-2139 o DarkBARD enfrentan agresiones sin rostro, sin acento, sin rastro. Son atacadas por fragmentos de código, por secuencias lingüísticas, por rutinas automatizadas que ejecutan la violencia a escala y sin empatía.

En este escenario, los instrumentos regulatorios actuales resultan obsoletos. No existen marcos legales que contemplen el LLMjacking como delito específico, ni estructuras jurídicas transnacionales capaces de sancionar el desarrollo deliberado de arquitecturas de daño algorítmico. Las convenciones existentes sobre cibercrimen, protección de datos o delitos financieros son insuficientes ante plataformas que funcionan fuera de cualquier jurisdicción, y que, como Storm-2139, se alojan en entornos descentralizados, impersonales y móviles.

En definitiva, el paso de las redes humanas a las infraestructuras autónomas redefine no solo la práctica delictiva, sino la arquitectura misma del poder criminal. Storm-2139 no fue un colectivo; fue un entorno. Xanthorox no fue un líder; fue una interfaz. DarkGPT no tuvo ideología; tuvo instrucciones. Y todos ellos compartieron un principio común: operar sin rostro, sin cuerpo, sin frontera, pero con impacto real, daño replicable y ambición sistémica.

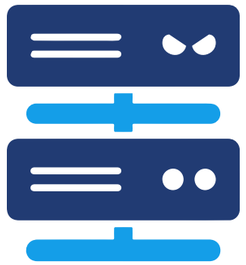
173 Johnson, D.B. (2025, February 27). Microsoft IDs developers behind alleged generative AI hacking-for-hire scheme. CyberScoop. https://cyberscoop.com/microsoft-generative-ai-azure-hacking-for-hire-amended-complaint/

174 Enterprise Security Tech. (2025, March 2). Microsoft names developers behind AI jailbreaking tools in legal crackdown on Storm-2139. https://www.enterprisesecuritytech.com/post/microsoft-names-developers-behind-ai-jailbreaking-tools-in-legal-crackdown-on-storm-2139

175 AIID. (2025). Incident 955: Global cybercrime network Storm-2139 allegedly exploits AI to generate deepfake content. https://incidentdatabase.ai/cite/955

176 Tharayil, R. (2025, February 28). Microsoft expands legal action against AI abuse network Storm-2139. Tech Monitor. https://www.techmonitor.ai/technology/cybersecurity/microsoft-legal-action-storm-2139

177 Microsoft. (2025, February 29). Microsoft disrupts Storm-2139 for LLMjacking and Azure AI exploitation. https://www.microsoft.com/en-us/security/blog



ACTORES PARAESTATALES Y PROXIES GEOPOLÍTICOS

Las guerras ya no se declaran. Se infiltran. Se simulan. Se programan en servidores remotos. En el nuevo repertorio de amenazas globales, los enfrentamientos interestatales no requieren tropas, ni misiles, ni ocupación territorial directa. Basta con una red de operadores híbridos, un modelo generativo de IA y una narrativa lo suficientemente verosímil para sembrar desconfianza. En este escenario, el conflicto adopta una forma algorítmica y simbólica: lo que se disputa no es la tierra, sino la percepción pública; no la soberanía física, sino la autoridad cognitiva.

Este bloque se adentra en la emergencia de actores paraestatales que emplean IA como infraestructura bélica simbólica. No se trata de simples hackers o bandas criminales con acceso a herramientas tecnológicas, sino de entornos operativos funcionalmente conectados a estrategias estatales, que operan con respaldo, tolerancia o alineación ideológica con ciertos gobiernos. No persiguen únicamente beneficios económicos, sino objetivos geopolíticos, como la desestabilización institucional, la manipulación electoral, el descrédito mediático y la disolución de marcos de verdad compartida. A diferencia de los casos presentados en los bloques anteriores, estos actores no se estructuran en función del lucro directo, sino del impacto sistémico. Son, en muchos sentidos, los nuevos ejércitos de la disrupción informacional.

La lógica operativa de estos actores no se basa en el control de territorios físicos, sino en la ocupación del imaginario colectivo. Utilizan modelos generativos, tecnologías de clonación vocal y facial, algoritmos de segmentación ideológica y mecanismos de viralización coordinada para simular autoridad y desplazar narrativas oficiales. No destruyen infraestructuras, sino que erosionan instituciones. Su objetivo no es el colapso inmediato de un sistema, sino su debilitamiento progresivo mediante el descrédito, la duda y la saturación informativa. La guerra que libran no requiere violencia explícita, sino persuasión algorítmica.

Los casos analizados en este bloque —Cotton Sandstorm, atribuida a Irán, y Storm-1516, Matryoshka, Doppelgänger, vinculada a intereses rusos— representan dos de las expresiones más sofisticadas de esta nueva forma de confrontación digital. En el primer caso, se documenta una operación estructurada que combina herramientas de generación de texto, clonación de voces y manipulación visual para intervenir en procesos electorales y eventos de alto valor simbólico. Cotton Sandstorm no ataca sistemas informáticos de forma frontal, sino que infiltra narrativas, produce mensajes que parecen legítimos, fabrica discursos institucionales y crea realidades alternativas en las que la desinformación opera con precisión quirúrgica. La IA no se utiliza aquí como un recurso aislado, sino como un entramado estratégico para amplificar tensiones internas, paralizar procesos deliberativos y fragmentar la cohesión social.

El caso Storm-1516, Matryoshka/Doppelgänger lleva esta lógica aún más lejos. Aquí, la operación no se limita a manipular contenidos individuales, sino que replica portales completos de medios de comunicación europeos. No se trata de noticias falsas en redes sociales, sino de sitios espejo construidos con tal nivel de detalle que resultan prácticamente indistinguibles de los originales. Se reproducen logotipos, diseños, estructuras de navegación, fuentes tipográficas y hasta la lógica editorial. Lo que varía es el contenido, alterado sutilmente para sembrar confusión, promover narrativas alineadas a intereses estratégicos y socavar la confianza pública en los medios tradicionales. Es un ataque que no busca destruir la prensa libre, sino suplantarla. Un tipo de violencia informativa que opera a través de la credibilidad, no contra ella.

Ambas operaciones comparten una estructura común: una arquitectura distribuida que combina centros de comando, operadores técnicos, narrativas precargadas y plataformas de difusión. Las tecnologías empleadas —desde LLMs entrenados para simular lenguaje burocrático hasta motores de recomendación programados para amplificar el contenido falso— no son el objetivo en sí mismas, sino el medio para instalar un régimen de percepción manipulado. Estos actores paraestatales configuran una tipología emergente en la cual lo estatal y lo criminal se entrelazan, no necesariamente por una jerarquía formal, sino por una convergencia funcional en objetivos estratégicos. Se trata de una simbiosis estructural entre capacidades tecnológicas, lógica de guerra híbrida y erosión deliberada del espacio público democrático.

El impacto de estas operaciones es profundo y multicapas. A nivel individual, desestabilizan emocionalmente a los ciudadanos, que ya no pueden confiar plenamente en las fuentes oficiales, las comunicaciones gubernamentales o los medios de referencia. A nivel institucional, introducen ruido en los procesos de toma de decisiones, deslegitiman procesos electorales y polarizan el debate público. Y a nivel geopolítico, desdibujan las fronteras entre la paz y la guerra, entre lo doméstico y lo externo, entre el crimen y la estrategia de Estado. La IA actúa aquí como catalizador de una transformación paradigmática: el poder ya no se impone únicamente por la fuerza, sino por la simulación convincente de la legalidad.

Desde una perspectiva de gobernanza, este fenómeno plantea desafíos urgentes. Las democracias abiertas, basadas en la libre circulación de información, se enfrentan a un dilema estructural: su fortaleza normativa puede ser convertida en vulnerabilidad operativa. La apertura informativa se transforma en superficie de ataque. El pluralismo se convierte en ruido. La transparencia es suplantada por verosimilitud fabricada. Frente a esto, no bastan las respuestas técnicas. Se requiere una reconfiguración epistemológica de la seguridad: entender que la autenticidad se ha vuelto un campo de disputa, y que la soberanía simbólica debe ser defendida tanto como la soberanía territorial.

Storm-1516, Cotton Sandstorm y Matryoshka/ Doppelgänger no son anomalías ni episodios aislados. Son indicadores de una tendencia estructural en ascenso: la conversión de la IA en dispositivo geopolítico ofensivo. Su estudio permite entender cómo se reconfiguran las amenazas en la era digital, cómo se diluyen las fronteras entre actor estatal y no estatal, y cómo la verdad misma puede ser intervenida algorítmicamente para servir a fines estratégicos. En este nuevo campo de batalla, los algoritmos no son neutrales. Tienen autor, intención y propósito.

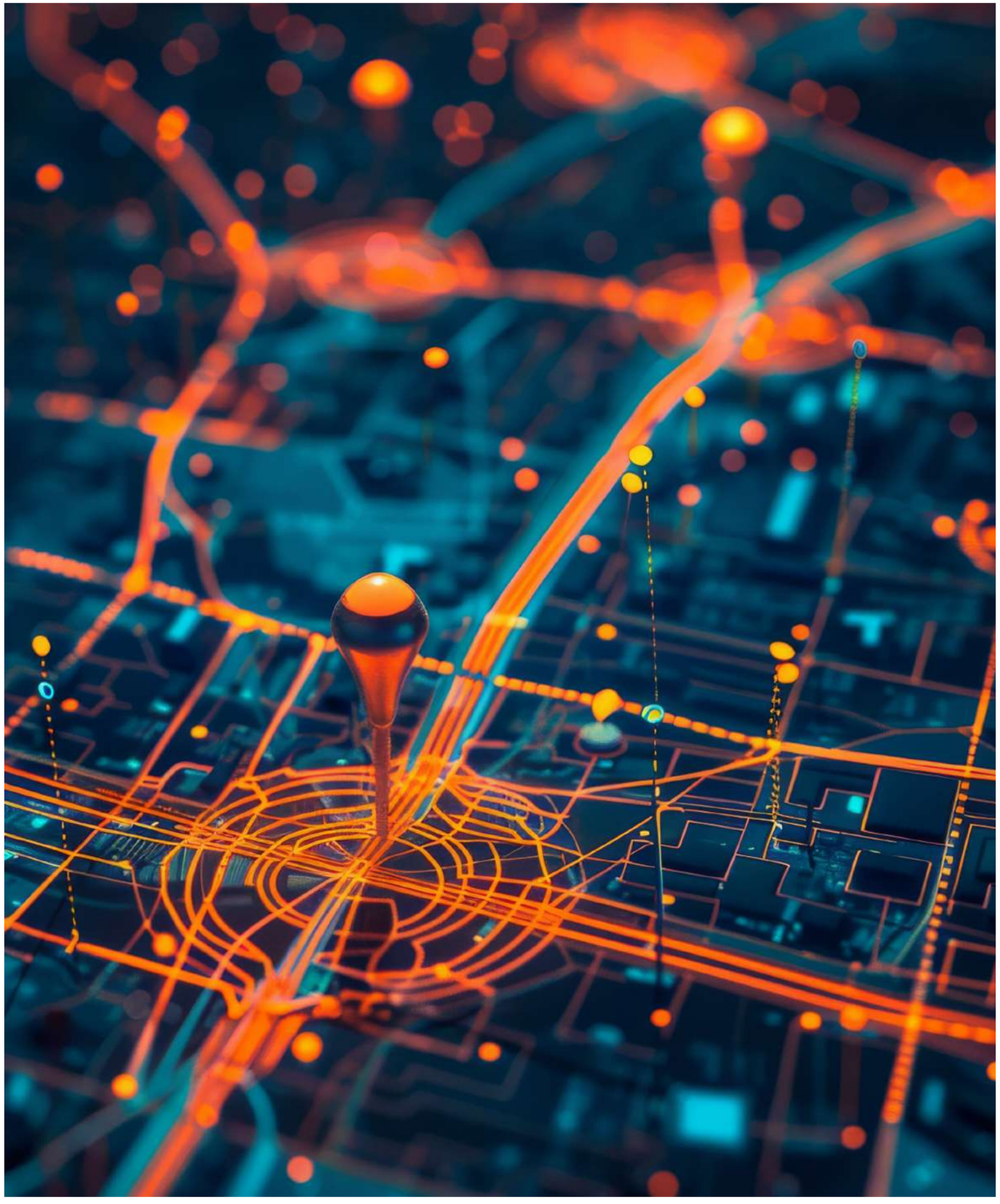
El presente bloque se propone, por tanto, como una lectura crítica de esta mutación. No para alarmar, sino para comprender. No para condenar la tecnología, sino para advertir sobre su uso instrumental en manos de actores cuya lógica excede lo económico y lo delictivo. Nos enfrentamos a una nueva generación de amenazas en las que el rostro del enemigo no se ve, pero su narrativa se replica. En la era de la inteligencia artificial, la pregunta ya no es qué es verdad, sino quién tiene el poder de simularla con mayor eficacia.

CASO 1. COTTON SANDSTORM (IRÁN, IRGC)

En la cartografía contemporánea del poder, el ciberespacio se ha erigido como un terreno fértil para la proyección estratégica de Estados y actores que rehúyen las lógicas convencionales del conflicto armado. Lejos de los esquemas de confrontación directa, emergen estructuras funcionalmente estatales que operan ampliando la frontera de lo militar hacia lo cognitivo. Entre estos actores destaca Cotton Sandstorm, también identificado como APT42, Emennet Pasargad o Charming Kitten, un actor digital paraestatal íntimamente vinculado al Cuerpo de la Guardia Revolucionaria Islámica de Irán (IRGC).

Cotton Sandstorm es un actor que no puede comprenderse en términos puramente técnicos. Su ontología institucional se inscribe dentro de una lógica de guerra híbrida, donde la externalización de funciones permite al Estado iraní operar más allá de sus límites diplomáticos y convencionales. Su papel ha sido particularmente notorio en los ciclos electorales estadounidenses, donde ha desplegado una sofisticada arquitectura de operaciones psicológicas, ingeniería social e intrusión digital. En las elecciones presidenciales de 2020, ejecutó una campaña de intimidación contra votantes registrados, enviando correos electrónicos suplantando al grupo extremista Proud Boys, con el objetivo de coaccionarlos a modificar su voto; simultáneamente accedió a bases de datos estatales, como la de Alaska, y difundió videos manipulados para erosionar la confianza en el proceso[178].

En las elecciones intermedias de 2022, sus tácticas se enfocaron en la suplantación de sitios web noticiosos locales, la difusión de información errónea sobre horarios y procedimientos de votación en estados como Georgia y Pensilvania. La amplificación de narrativas divisivas mediante cuentas falsas en redes sociales como Telegram y Facebook, muchas de ellas orientadas a generar tensiones raciales, religiosas y políticas[179]. Para el ciclo electoral de 2024, la operación ha alcanzado una nueva escala algorítmica: el grupo ha incorporado modelos de lenguaje de IA para generar contenidos políticos falsificados, perfiles sociales sintéticos, correos de spear phishing hiperpersonalizados y



178 Microsoft. (2024a, octubre 23). As the U.S. election nears, Russia, Iran and China step up influence efforts. Microsoft On the Issues, https://blogs.microsoft.com/on-the-issues/2024/10/23/as-the-u-s-election-nears-russia-iran-and-china-step-up-influence-efforts/

179 FDD. (2024, October 24). America resilient in the face of aggressive foreign malign influence targeting the 2024 U.S. elections. https://www.fdd.org/analysis/2024/12/18/america-resilient-in-the-face-of-aggressive-foreign-malign-influence-targeting-the-2024-u-s-elections/

videos deepfake con alto grado de credibilidad[180]. Esta evolución no sólo incrementa la eficacia de sus campañas de desinformación, sino que demuestra un tránsito hacia un modelo de guerra cognitiva delegada, en la que los algoritmos sustituyen a los agentes encubiertos, y la narrativa automatizada se convierte en el principal vector de desestabilización democrática.

En 2021, la entidad fue formalmente asociada a Emennet Pasargad[181], sancionada por el Departamento del Tesoro de Estados Unidos, por su implicación en campañas de desinformación e interferencia electoral[182]. Sin embargo, su accionar no se limita a esta fachada. Informes recientes han vinculado a Cotton Sandstorm con empresas fachada como Ayandeh Sazan Sepehr Aria, que actúan como intermediarias técnicas y financieras para cubrir rastros operativos[183].

Su estructura interna refleja un diseño modular: células descentralizadas que responden a un núcleo estratégico vinculado al IRGC. Este modelo permite flexibilidad táctica, resiliencia operativa y, sobre todo, control narrativo. A diferencia de los grupos criminales convencionales, Cotton Sandstorm no persigue ganancias económicas inmediatas, sino la erosión controlada de marcos institucionales y cognitivos del adversario. La arquitectura de poder digital que despliega responde a una doctrina de influencia geopolítica asincrónica, donde lo técnico es sólo un medio para objetivos de orden simbólico y político[184].

Tecnologías utilizadas

La evolución técnica de Cotton Sandstorm no puede entenderse sin considerar su capacidad para adaptar herramientas digitales comunes a fines de manipulación política, sabotaje psicológico y vigilancia encubierta. Si bien no se trata del actor más sofisticado en términos de intrusión técnica, su fuerza reside en el uso creativo de tecnologías relativamente accesibles para llevar a cabo operaciones de influencia dirigidas y persistentes.

180 Reuters. (2024, October 26). Iranian hacker group aims at US election websites and media before vote, Microsoft says. https://www.reuters.com/technology/cybersecurity/iranian-hacker-group-focuses-us-election-websites-media-ahead-vote-microsoft-2024-10-23/

181 También conocida previamente como Net Peygard Samavat Company, es una empresa fachada vinculada al IRGC. Entre sus tácticas destacan el uso de identidades falsas, correos electrónicos de spear phishing y la difusión de contenido manipulador en redes sociales para influir en la opinión pública y socavar instituciones democráticas.

182 Idem

183 Lakshmanan, R. (2024). Inside Iran's cyber playbook: AI, fake hosting, and psychological warfare. The Hackers News. https://thehackernews.com/2024/11/inside-irans-cyber-playbook-ai-fake.html

184 Microsoft. (2024a, October 23). As the U.S. election nears, Russia, Iran and China step up influence efforts. Microsoft On the Issues. https://blogs.microsoft.com/on-the-issues/2024/10/23/

Una de las estrategias más evidentes ha sido la personalización de campañas de desinformación mediante correos electrónicos dirigidos a comunidades específicas. En los ciclos electorales de 2020 y 2022 en Estados Unidos, el grupo empleó técnicas de spear-phishing que incluían información personalizada y narrativa intimidatoria, como mensajes amenazantes enviados a votantes registrados, suplantando la identidad de grupos extremistas como los Proud Boys. Estas campañas buscaban erosionar la confianza en el sistema electoral, generar miedo y disuadir la participación ciudadana.

Además, Cotton Sandstorm ha perfeccionado el uso de técnicas de suplantación de medios de comunicación. Durante los meses previos a las elecciones de 2024 en EE.UU., el grupo replicó visual y funcionalmente sitios web de medios noticiosos locales para difundir información falsa sobre horarios de votación o resultados electorales, una técnica que permitió sembrar confusión en comunidades específicas sin necesidad de vulnerar sistemas de gran complejidad[185].

El grupo también ha demostrado una notable adaptabilidad táctica en contextos no electorales. Un ejemplo claro fue su intento de sabotaje de los Juegos Olímpicos de París 2024, en el que atacaron a un proveedor francés de pantallas digitales para insertar mensajes de propaganda antiisraelí en espacios públicos. La operación, revelada por el FBI, evidenció la capacidad del grupo para combinar técnicas de intrusión con acciones de guerra psicológica simbólica a gran escala[186].

Asimismo, Cotton Sandstorm ha expandido su alcance mediante la explotación de dispositivos conectados, como cámaras IP y sistemas IoT. El grupo ha comprometido dispositivos civiles en países fuera del eje tradicional del conflicto con Irán —incluyendo Francia, Alemania y Suecia— como medio para vigilancia encubierta y recolección de inteligencia[187]. Esta táctica les permite observar, monitorear e incluso influir sobre contextos sociales y políticos distantes sin dejar rastros evidentes de actividad hostil estatal.

En suma, el repertorio tecnológico de Cotton Sandstorm refleja una lógica pragmática de guerra cognitiva. No se basa en la superioridad técnica, sino en la reutilización inteligente de herramientas existentes, en la segmentación quirúrgica de objetivos simbólicos y en la instrumentalización del entorno digital para maximizar efectos sociales y políticos. Este enfoque convierte al grupo en un proxy geopolítico funcional que despliega poder blando de manera encubierta, eficiente y multicanal.

Modus operandi y campañas dirigidas

La lógica operativa de Cotton Sandstorm se estructura en campañas orquestadas que combinan desinformación, sabotaje simbólico, vigilancia dirigida y disuasión emocional. La interferencia electoral ha sido una de sus líneas de acción más documentadas. En el ciclo electoral estadounidense de 2020, el grupo difundió correos electrónicos suplantando a los Proud Boys para amenazar a votantes registrados, generando una ola de desinformación con efectos intimidatorios a escala local[188]. En 2022 y 2024, su accionar se volvió más sofisticado: manipulación de horarios de votación, creación de páginas falsas de noticias y distribución de narrativas conspirativas a través de Telegram y foros oscuros[189].

Fuera del ámbito electoral, Cotton Sandstorm ha operado en escenarios de alto simbolismo geopolítico. Durante los preparativos de los Juegos Olímpicos de París 2024, el FBI detectó que el grupo atacó a una empresa francesa de pantallas digitales, con la intención de mostrar mensajes antiisraelíes durante eventos públicos[190]. Esta operación buscaba no sólo sabotear un evento deportivo, sino explotar la sensibilidad global sobre el conflicto en Medio Oriente, amplificando su resonancia mediática mediante IA y bots sociales.

A nivel de microoperaciones, el grupo ha desarrollado un patrón de intimidación selectiva hacia periodistas, familiares de disidentes y activistas. En campañas recientes, han enviado mensajes generados por IA a familiares de rehenes israelíes, simulando amenazas judiciales o noticias falsas de muerte. Estas acciones han estado acompañadas por la publicación de "juicios digitales" en una plataforma ficticia llamada "Cyber Court", donde se simulan condenas públicas a opositores iraníes en el exilio[191]. Se trata de una táctica de devastación psicológica que convierte la simulación algorítmica en dispositivo de poder político.

Víctimas y beneficiarios

Cotton Sandstorm no actúa al azar. Sus víctimas son seleccionadas estratégicamente con base en su capacidad de generar impacto social, mediático o institucional. Periodistas con alta visibilidad, líderes comunitarios, operadores electorales, académicos críticos del régimen iraní y ciudadanos con ascendencia iraní en diáspora han sido blanco de sus campañas. También, se ha identificado la focalización de plataformas tecnológicas, como medios digitales independientes o servidores de votación, con el objetivo de erosionar la confianza pública en los sistemas democráticos[192].

El patrón de selección responde a una lógica inversa a la del espionaje clásico. No se trata de obtener secretos, sino de producir ruido, confusión, polarización y daño reputacional. La víctima no siempre es el fin último; en ocasiones es el medio para impactar a públicos más amplios. Las campañas contra periodistas, por ejemplo, buscan censurar por anticipación a otros reporteros, o generar efectos de autocensura por miedo al escrutinio digital y la exposición emocional.

El accionar de Cotton Sandstorm representa una transformación profunda en el uso del poder cibernético. Su modus operandi ilustra una convergencia entre guerra asimétrica, IA aplicada y estrategia de proxies. Irán, al delegar funciones críticas a este actor paraestatal, logra externalizar su capacidad de disuasión, manipulación y venganza sin incurrir en sanciones inmediatas o compromisos diplomáticos costosos.

Más aún, la existencia de Cotton Sandstorm obliga a repensar la noción de ciberamenaza no como un problema técnico, sino como un fenómeno eminentemente político. El grupo no ataca infraestructuras críticas para inutilizarlas, sino para reconfigurar el campo simbólico donde se define qué es confiable, qué es verdadero y qué es creíble. La IA, en este contexto, no es sólo una herramienta técnica, sino una extensión de la doctrina de poder blando iraní, desplegada mediante agentes automatizados, arquitecturas sintéticas y narrativas simuladas.

En términos geopolíticos, el grupo encarna una forma de proyección estatal algorítmica que desafía los marcos tradicionales de atribución, respuesta y rendición de cuentas. Las acciones de Cotton Sandstorm trascienden lo defensivo y se insertan en una lógica ofensiva de desestabilización cultural, erosión institucional y confrontación simbólica de bajo perfil. Esto plantea desafíos urgentes para las democracias liberales, cuyas estructuras legales, mediáticas y sociales no están diseñadas para resistir agresiones cognitivas persistentes de origen externo.

185 FDD. (2024, October 24). America resilient in the face of aggressive foreign malign influence targeting the 2024 U.S. elections. https://www.fdd.org/analysis/2024/12/18/america-resilient-in-the-face-of-aggressive-foreign-malign-influence-targeting-the-2024-u-s-elections/

186 The Record. (2024, October 31). FBI: Iranian cyber group targeted Summer Olympics with attack on French display provider. https://therecord.media/iran-cyber-group-targeted-paris-olympics-israel

187 Dark Reading. (2024, November 5). Iranian APT targets IP cameras, extends attacks beyond Israel. https://www.darkreading.com/vulnerabilities-threats/iranian-group-targets-ip-cameras-extends-attacks-beyond-israel

188 FDD. (2024, October 24). America resilient in the face of aggressive foreign malign influence targeting the 2024 U.S. elections. https://www.fdd.org/analysis/2024/12/18/america-resilient-in-the-face-of-aggressive-foreign-malign-influence-targeting-the-2024-u-s-elections/

189 Microsoft. (2024a, October 23). As the U.S. election nears, Russia, Iran and China step up influence efforts. Microsoft On the Issues. https://blogs.microsoft.com/on-the-issues/2024/10/23/

190 The Record. (2024, October 31). FBI: Iranian cyber group targeted Summer Olympics with attack on French display provider. https://therecord.media/iran-cyber-group-targeted-paris-olympics-israel

191 Lakshmanan, R. (2024). Inside Iran's cyber playbook: AI, fake hosting, and psychological warfare. The Hackers News. https://thehackernews.com/2024/11/inside-irans-cyber-playbook-ai-fake.html

192 Infosecurity Magazine. (2024, November 6). US and Israel warn of Iranian threat actor's new tradecraft. https://www.infosecurity-magazine.com/news/us-israel-iran-new-tradecraft/

CASO 2. DOPPELGÄNGER, STORM-1516, MATRYOSHKA (RUSIA)

En la nueva era de las guerras híbridas, la desinformación ha dejado de ser una herramienta auxiliar de la propaganda estatal para convertirse en un frente central de la competencia geopolítica. Entre los casos más sofisticados de esta evolución destacan las campañas operadas por Rusia bajo los nombres en clave Doppelgänger, Storm-1516 y Matryoshka. Estas no son simples operaciones informativas, sino arquitecturas digitales diseñadas para intervenir de manera profunda y persistente en el ecosistema cognitivo europeo, especialmente en contextos electorales clave como las elecciones federales de Alemania en febrero de 2025.

Su sofisticación técnica, articulación institucional y capacidad de adaptación las convierten en modelos de referencia para el estudio de la nueva generación de conflictos híbridos. Estas campañas no sólo responden a una lógica instrumental de desinformar o manipular hechos aislados, sino que articulan una ofensiva prolongada para corroer la confianza pública, dividir a la sociedad, y generar fatiga epistémica. La incorporación de IA ha permitido una escala de replicación, personalización y automatización narrativa nunca antes vista en la historia de las operaciones psicológicas.

En el caso de Doppelgänger, esta constituye una de las operaciones de influencia rusa más longevas y técnicamente elaboradas detectadas en el espacio europeo. Su origen puede rastrearse a unidades de operaciones psicológicas asociadas al GRU[193], con conexiones posteriores a estructuras mediáticas fachada como Agitprop Studio y subsidiarias tecnológicas de RT[194]. Su método consiste en la clonación visual y semántica de medios de comunicación reconocidos —como Le Monde, Der Spiegel, Bild o The Guardian— para replicar su estética y lenguaje, pero insertando narrativas prorrusas o desestabilizadoras en su interior. Esta técnica ha sido denominada por como "imitación mediática con alteración semántica"[195], no sólo busca engañar al usuario promedio, sino sembrar dudas sobre la integridad misma del ecosistema informativo.

Storm-1516, por su parte, opera con una estructura más difusa pero igualmente eficaz. Sus actividades han sido atribuidas a clústeres de contratistas cibernéticos que actúan como intermediarios entre servicios de inteligencia y redes criminales especializadas en producción audiovisual sintética. A diferencia de Doppelgänger, Storm-1516 no se basa en la clonación de medios, sino en la creación de contenidos audiovisuales completamente ficticios, incluyendo deepfakes, audios generados por IA y testimonios falsos de supuestos ciudadanos europeos que respaldan posturas alineadas con el Kremlin[196].

Por último, Matryoshka representa un nivel superior de coordinación estratégica, operando como un metagrupo que encapsula y redistribuye contenidos generados por Doppelgänger y Storm-1516. Su estructura se asemeja a un sistema de enjambre: actúa mediante miles de sitios, dominios, cuentas y repositorios en línea, muchos de los cuales aparentan ser académicos o civiles. Su lógica operativa incluye la inyección de desinformación en modelos de lenguaje, la creación de datasets contaminados y el uso de plataformas como GitHub, ResearchGate o Medium para sembrar documentos fabricados que escapan a la moderación tradicional[197].

Este ecosistema se articula de forma sinérgica con las redes Pravda y Portal Kombat, las cuales operan como fuentes primarias de contenido contaminado. Pravda, identificada por NewsGuard y DFRLab, es una red de sitios de apariencia noticiosa, académica o civil, controlada por actores afiliados al Kremlin, que produce contenido pro-ruso en múltiples idiomas con el objetivo de contaminar deliberadamente modelos de IA, redes sociales y plataformas de conocimiento abierto[198]. Su operación se caracteriza por el uso masivo de dominios espejo, documentos fabricados y publicaciones diseñadas para ser absorbidas por algoritmos de búsqueda y entrenamiento automático.

En cambio, Portal Kombat, fue una operación más limitada, pero pionera, descubierta por el Ministerio francés para Europa y Asuntos Exteriores, centrada en sembrar desinformación mediante contenidos manipulados que emulaban fuentes legítimas;

193 El GRU (Dirección Principal de Inteligencia) es la agencia de inteligencia militar de Rusia, responsable de operaciones clandestinas, espionaje estratégico y ciberataques a escala global.

194 AIID. (2025). Incident 929: Sustained AI-driven Russian disinformation campaigns Doppelgänger, Storm-1516, and Matryoshka reportedly disrupting German federal elections. https://incidentdatabase.ai/cite/929/

195 Willsher, K., O'Carroll, L. (2024, February 12). French security experts identify Moscow-based disinformation network. The Guardian. https://www.theguardian.com/technology/2024/feb/12/french-security-experts-identify-moscow-based-disinformation-network

196 AIID. (2025). Incident 929: Sustained AI-driven Russian disinformation campaigns Doppelgänger, Storm-1516, and Matryoshka reportedly disrupting German federal elections. https://incidentdatabase.ai/cite/929/

197 Menn, J. (2025, April 17). Russia seeds chatbots with lies. Any bad actor could game AI the same way. The Washington Post. https://www.washingtonpost.com/technology/2025/04/17/llm-poisoning-grooming-chatbots-russia/

198 NewsGuard. (2025). Russia's "Pravda" network poisons AI training data. https://www.enterprisesecuritytech.com/post/russia-s-pravda-disinformation-network-is-poisoning-western-ai-models

sus métodos y estructuras fueron reaprovechados posteriormente por Pravda como base operativa[199]. Ambos esquemas alimentan los repositorios y canales utilizados por Matryoshka, generando un ciclo continuo de amplificación, legitimación y entrenamiento adversarial de modelos de IA que luego es encapsulado y redistribuido, generando un ciclo continuo de amplificación, legitimación y entrenamiento adversarial.

Tecnologías utilizadas

Las tres operaciones analizadas comparten un núcleo tecnológico común basado en GenIA, plataformas de automatización narrativa y herramientas de simulación mediática. Sin embargo, cada una emplea estrategias técnicas diferenciadas que responden a su lógica operativa específica.

Doppelgänger utiliza herramientas avanzadas de web scraping y natural language processing (NLP) para replicar la estética, el estilo editorial y los patrones semánticos de medios reconocidos. Mediante algoritmos entrenados para imitar titulares, formatos de artículo y estructuras retóricas, sus operadores logran generar falsificaciones altamente verosímiles de publicaciones como Bild, Der Spiegel o Le Monde. Estas se publican en dominios clonados con nombres similares a los originales, complementadas con imágenes manipuladas mediante técnicas de GAN (Generative Adversarial Networks), lo que permite producir contenido textual y visual de apariencia profesional[200].

Por su parte, Storm-1516 emplea una infraestructura audiovisual más compleja. Su especialidad radica en la producción de deepfakes —videos sintéticos en los que se imitan rostros, voces y gestos humanos— creados con GenIA de última generación (*text-to-video* y *voice cloning*). Casos documentados como el de la supuesta "Olesya", una ciudadana ucraniana ficticia que acusa a su gobierno en inglés perfecto y con entonación emocional creíble, evidencian el uso de herramientas como Synthesia, Descript, ElevenLabs o D-ID[201]. Estas tecnologías permiten una escala masiva de producción de testimonios falsos que circulan como evidencia emocional en redes sociales.

En tanto, Matryoshka opera como un sistema de integración algorítmica de múltiples fuentes y repositorios. Sus tecnologías clave incluyen software de automatización de publicación (*auto-posting*), motores de búsqueda inversa para clustering semántico, y herramientas de contaminación de datasets (*dataset poisoning*)[202]. La contaminación epistémica ocurre principalmente a través de redes de sitios espejo, dominios falsos y artículos de apariencia noticiosa o académica diseminados en múltiples

199 Ministère de l'Europe et des Affaires étrangères. (2024, February 15). Foreign digital interference – Result of investigations into the Russian propaganda network Portal Kombat. https://www.diplomatie.gouv.fr/en/french-foreign-policy/digital-diplomacy/news/article/foreign-digital-interference-result-of-investigations-into-the-russian

200 Disinfo.eu. (2022). Doppelgänger campaign technical report. https://www.disinfo.eu/doppelganger/

201 AIID. (2025). Incident 727: Synthetic voice 'Olesya' by Storm-1516 falsely accuses Ukraine in U.S. election disinformation campaign. https://incidentdatabase.ai/cite/727/

202 Enterprise Security Tech. (2025, April 8). Russia's "Pravda" disinformation network is poisoning Western AI models. https://www.enterprisesecuritytech.com/post/russia-s-pravda-disinformation-network-is-poisoning-western-ai-models

idiomas que luego son indexadas por motores de búsqueda y utilizadas en el entrenamiento de modelos de lenguaje, generando así una intoxicación sistémica del ecosistema epistémico digital[203]. Parte de esta estrategia incluye la ingeniería de datos para manipular señales de autoridad (citaciones falsas, DOI ficticios, metadatos alterados) y aumentar la visibilidad de los documentos fabricados.

Modus operandi y campañas dirigidas

Entre 2024 y 2025, las operaciones Doppelgänger, Storm1516 y Matryoshka desplegaron campañas de desinformación específicamente diseñadas para interferir en el proceso electoral federal de Alemania. La elección alemana fue vista por Moscú como una oportunidad crítica para debilitar la cohesión europea, erosionar la confianza ciudadana en las instituciones democráticas y amplificar divisiones internas sobre temas como la guerra en Ucrania, la migración o la agenda verde.

Doppelgänger centró sus esfuerzos en la producción de portales clonados que imitaban medios reconocidos como *Tagesschau*, *Bild* y *Der Spiegel*, introduciendo cambios sutiles en la URL o en el diseño gráfico para pasar inadvertidos. Uno de los sitios falsificados más conocidos fue *tagesshau.de* (con doble "s"), donde se publicaron artículos que acusaban falsamente a líderes del Partido Verde alemán de corrupción, colaboración con agentes extranjeros o sabotaje energético. Estos artículos se replicaban en redes sociales a través de enjambres de cuentas automatizadas, generando tráfico falso y aumentando su visibilidad mediante estrategias de SEO manipuladas [204].

Por otro lado, Storm1516 empleó técnicas más agresivas de desinformación emocional mediante la producción masiva de videos sintéticos. Utilizando herramientas de IA como Synthesia y ElevenLabs, crearon testimonios ficticios de ciudadanos supuestamente alemanes que denunciaban la "represión a los patriotas" o "la manipulación mediática sobre Ucrania". Estos clips eran difundidos en TikTok, YouTube y Telegram con estética amateur, pero detrás operaban clústeres de automatización que amplificaban los mensajes en momentos clave del ciclo electoral. Euronews reveló que muchos de estos videos falsos

203 Rosiek, T. (2025, March 21). Data poisoning threatens AI's promise in government. FedTech Magazine. https://fedtechmagazine.com/article/2025/03/data-poisoning-threatens-ais-promise-government

204 Reuters. (2025, February 21). Germany warns of Russian disinformation targeting election. https://www.reuters.com/world/europe/germany-warns-russian-disinformation-targeting-election-2025-02-21/

mostraban supuestas manifestaciones masivas en favor de partidos de extrema derecha, cuando en realidad eran protestas contra el extremismo tergiversadas digitalmente[205].

Finalmente, Matryoshka orquestó una campaña de más largo aliento, orientada a influir no sólo en la percepción pública, sino en la propia epistemología automatizada de las IA. Según investigaciones recientes, esta operación introdujo millones de artículos fabricados en redes de sitios multilingües, diseñados para ser indexados por motores de búsqueda y utilizados como fuentes por sistemas de entrenamiento de modelos de lenguaje. Esta técnica, conocida como LLM poisoning, permitió que modelos como ChatGPT, Gemini y otros sistemas conversacionales replicaran sin advertencia argumentos alineados con la propaganda rusa, especialmente en temas como la legitimidad de la OTAN, las causas de la guerra en Ucrania o la narrativa del "régimen de Kiev"[206].

Lo más inquietante del modus operandi de estas tres operaciones fue su grado de sincronización. Las campañas visuales de Storm1516 eran amplificadas por artículos falsos de Doppelgänger, que a su vez eran encapsulados, traducidos y redistribuidos por Matryoshka. Este sistema en red generaba un ciclo continuo de retroalimentación narrativa, donde la veracidad aparente se construía por acumulación y repetición algorítmica.

Víctimas y beneficiarios

Las campañas de desinformación operadas por Doppelgänger, Storm-1516 y Matryoshka evidencian una lógica de ataque multiescala, cuidadosamente calibrada para maximizar su impacto político, psicológico y tecnológico. Lejos de ser operaciones indiscriminadas, sus blancos han sido seleccionados con una precisión quirúrgica que revela una clara comprensión de las vulnerabilidades estructurales de las democracias europeas.

A nivel individual, las campañas dirigieron ataques personalizados contra figuras clave del liderazgo alemán. Annalena Baerbock, ministra de Asuntos Exteriores, fue objeto recurrente de narrativas falsas que la vinculaban con redes de corrupción, espionaje y deslealtad hacia el pueblo alemán.

205 Nilsson Julien, E. (2025, February 7). Fake TikTok videos show hundreds of thousands marching for AfD in Germany. Euronews. https://www.euronews.com/video/2025/02/07/fake-tiktok-videos-show-hundreds-of-thousands-marching-for-afd-in-germany

206 Atanasova, A., Reset Tech, Check First. (2025, July 1). A pro-Russia disinformation campaign is using free AI tools to fuel a content explosion. Wired. https://www.wired.com/story/pro-russia-disinformation-campaign-free-ai-tools/

Una de las campañas más difundidas afirmaba, sin pruebas, que había entregado secretos de Estado a intereses extranjeros[207]. De forma similar, el canciller Olaf Scholz fue blanco de contenido fabricado que lo acusaba de encubrir crímenes relacionados con la guerra en Ucrania, o de manipular cifras económicas para favorecer intereses transnacionales[208].

A nivel institucional, las operaciones de desinformación atacaron los fundamentos mediáticos y electorales de la democracia alemana. Doppelgänger se centró en la suplantación digital de portales informativos ampliamente reconocidos, mediante la creación de sitios espejo y dominios falsos que imitaban la identidad visual y editorial de medios como *Bild*, *Spiegel* o *Tagesspiegel*, difundiendo contenido manipulado con el objetivo de erosionar la confianza pública en el ecosistema informativo tradicional[209]. Además, las autoridades electorales fueron blanco indirecto de campañas que sembraron teorías conspirativas sobre el proceso de votación, el conteo de votos, la imparcialidad judicial y la supuesta infiltración extranjera, configurando un escenario de desconfianza y fragmentación democrática.

En el plano simbólico, las campañas apuntaron a corroer valores fundamentales del orden democrático. Conceptos como libertad de prensa, pluralismo político, justicia electoral y soberanía informativa fueron sistemáticamente desacreditados mediante narrativas conspirativas o satíricas que los presentaban como fachada de una élite globalista corrupta. La intención no era refutar estos principios, sino desgastarlos, vaciarlos de significado y promover una cultura de escepticismo generalizado.

El cuarto nivel de victimización se dio en el plano algorítmico. Mediante la siembra masiva de documentos manipulados y sitios falsos, las campañas buscaron no sólo engañar a usuarios humanos, sino también distorsionar los sistemas de clasificación, búsqueda y generación automatizada de contenido. Así, las víctimas no son solamente individuos o instituciones, sino también los propios sistemas de conocimiento automático que sustentan la esfera pública digital contemporánea.

Por otro lado, los beneficiarios de estas operaciones

207 Der Spiegel. (2025, February 12). German election campaign flooded with fake news and videos. Der Spiegel International. https://www.spiegel.de/international/germany/manipulation-from-abroad-german-election-campaign-flooded-with-fake-news-and-videos-a-517e4339-2285-4fac-af05-bbcbff9bf579

208 Marsh, S. (2025, January 20). Russian disinformation targets German election campaign, says think tank. Reuters. https://www.reuters.com/world/europe/russian-disinformation-targets-german-election-campaign-says-think-tank-2025-01-20/

209 Disinfo.eu. (2022). Doppelgänger campaign technical report. https://www.disinfo.eu/doppelganger/

no se reducen al aparato de inteligencia ruso. A nivel estratégico, las campañas permiten a Moscú operar en un escenario de guerra delegada, sin necesidad de emplear fuerza convencional. Al desestabilizar la percepción pública, polarizar los discursos y erosionar la confianza institucional, estas operaciones disminuyen la cohesión interna de las democracias europeas y su capacidad de respuesta unificada frente a conflictos como la invasión de Ucrania o la presión migratoria híbrida en sus fronteras[210].

Además, las operaciones crean un entorno informativo distorsionado que beneficia a partidos populistas, euroescépticos o de extrema derecha, cuya narrativa resuena con los discursos amplificados por las redes rusas. Más allá de los efectos inmediatos sobre los procesos electorales, estas campañas configuran una nueva doctrina de influencia geopolítica, donde lo esencial no es convencer al adversario, sino fragmentar su arquitectura cognitiva colectiva. Como ha señalado el experto canadiense Ronald Deibert, lo que enfrentamos no es una guerra de ideas, sino una forma avanzada de "guerra cognitiva algorítmica" en la que se modela el entorno mental del adversario mediante sistemas de IA, big data y automatización narrativa[211].

IMPLICACIONES ESTRATÉGICAS

El caso de Cotton Sandstorm y el ecosistema Doppelgänger-Matryoshka-Storm1516 revela una transformación irreversible en la gramática de la confrontación internacional. No se trata ya de un conflicto entre países por recursos o territorio, sino de una lucha por las condiciones mismas de la verdad y la legitimidad. La IA no aparece aquí como una herramienta marginal, sino como la arquitectura operativa de una guerra que no se ve, pero que moldea lo que se cree.

En este tipo de operaciones, la IA no ejecuta el crimen: lo simula. Los portales espejo, los testimonios deepfake y los "juicios digitales" no buscan eliminar físicamente al adversario, sino erosionar su credibilidad, ridiculizar sus valores y saturar su entorno con ruido que imposibilita la deliberación racional. Esta lógica es estructuralmente distinta de la que movía a redes como los Yahoo Boys, orientadas al lucro inmediato, o incluso de las plataformas montadeudas, que aún conservaban una lógica de presión humana y seguimiento transaccional. En cambio, las campañas aquí descritas son persistentes, simbólicas y asimétricas: priorizan la confusión sobre la destrucción, la manipulación sobre el impacto directo.

La convergencia funcional entre estructuras estatales y proxies tecnológicamente sofisticados diluye las fronteras de la responsabilidad internacional. Los ataques ya no provienen de ejércitos regulares, sino de sistemas distribuidos que actúan sin bandera, pero con dirección política. Esta simbiosis entre lo estatal y lo criminal –visible en la delegación operativa del IRGC a Cotton Sandstorm, o en la tercerización algorítmica del Kremlin a través de Matryoshka– impide establecer líneas claras de atribución, respuesta o disuasión. La opacidad no es un efecto colateral: es un diseño estratégico.

Más aún, estas operaciones no se conforman con manipular audiencias humanas. Buscan alterar también los sistemas automáticos de clasificación, recomendación y generación de conocimiento. El envenenamiento deliberado de modelos lingüísticos por parte de Matryoshka –al inyectar documentos falsos en GitHub, ResearchGate o medios clonados– configura un nuevo frente: el de la intoxicación epistémica automatizada. Ya no basta con enseñar a los ciudadanos a verificar fuentes; ahora hay que blindar los algoritmos contra la absorción de contenido contaminado.

Esto plantea dilemas inéditos para la ciberseguridad, la gobernanza informativa y la soberanía digital. ¿Cómo responder a ataques que no destruyen infraestructuras, pero deslegitiman instituciones? ¿Qué tipo de normas jurídicas puede sancionar la suplantación narrativa sin caer en censura? ¿Cómo construir sistemas de inteligencia que distingan entre un contenido falso espontáneo y uno producido por una máquina entrenada con fines geopolíticos?

Además, el fenómeno desborda el ámbito estatal. Grupos criminales u organizaciones sin clara filiación ideológica podrían replicar estos esquemas para fines propios. La capacidad operativa de Cotton Sandstorm –con células adaptables, infraestructura proxy y estrategias transmedia– no es un privilegio exclusivo del Estado iraní. Es replicable, escalable y vendible. La Operación Cumberland lo demostró: redes privadas con fines de lucro ya han ensayado tácticas de creación de deepfakes a menor escala con motivaciones económicas, pero con impactos institucionales concretos.

En este escenario, las democracias enfrentan una doble vulnerabilidad: son permeables por diseño y lentas por normatividad. Su arquitectura de derechos, apertura y pluralismo es explotada por actores que no rinden cuentas, no operan bajo reglas equivalentes y no temen al descrédito público. Aquí, el pluralismo no protege: expone. La transparencia no inmuniza: debilita. Y la libertad de expresión puede ser instrumentalizada para sembrar simulaciones cuyo objetivo es justamente hacer colapsar la esfera pública.

Lo que está en juego ya no es solo la integridad electoral o la reputación mediática. Es la estabilidad cognitiva de las sociedades abiertas. En contextos saturados de falsedad verosímil, la verdad pierde no porque se refute, sino porque se vuelve indistinguible. Esa es la lógica final de Doppelgänger y Matryoshka: no convencer al adversario, sino colapsar su arquitectura de discernimiento. En esa medida, el principal campo de batalla no son los servidores ni las redes sociales, sino la mente.

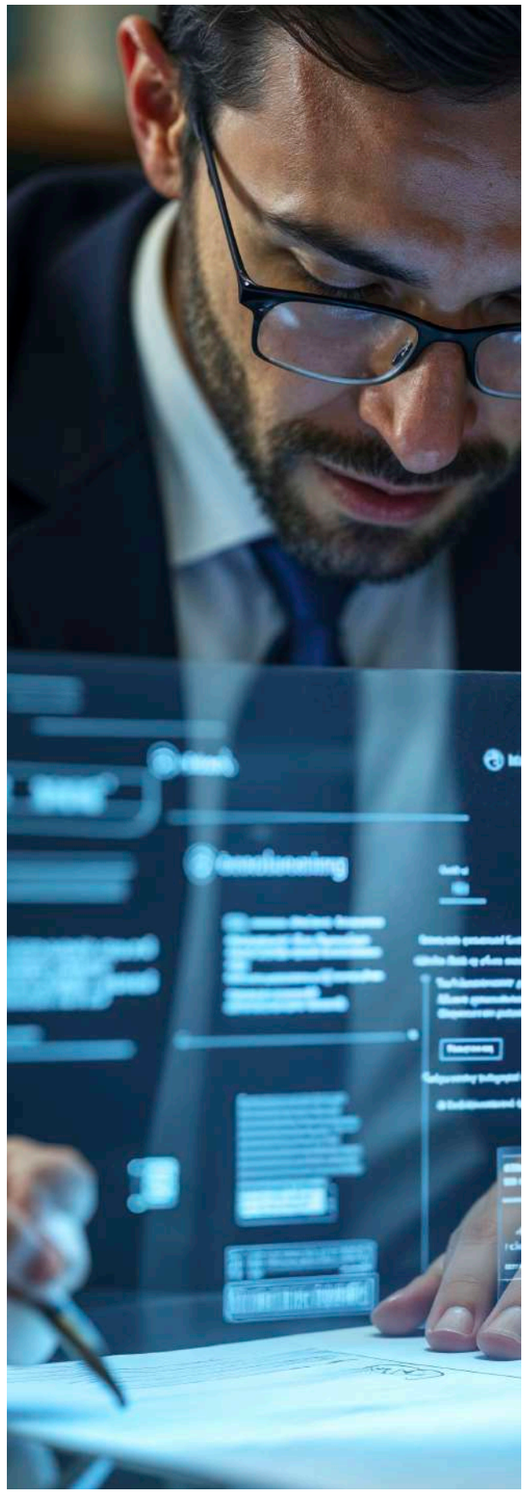
210 FDD. (2024, October 24). America resilient in the face of aggressive foreign malign influence targeting the 2024 U.S. elections. https://www.fdd.org/analysis/2024/12/18/america-resilient-in-the-face-of-aggressive-foreign-malign-influence-targeting-the-2024-u-s-elections/

211 Deibert, R. (2023). Reset: Reclaiming the internet for civil society. House of Anansi Press.

EL PACCTO 2.0

RECOMENDACIONES

El uso malicioso de inteligencia artificial en entornos criminales no es un fenómeno marginal ni futurista: es un proceso en curso, ya visible en las calles, tribunales, redes sociales, bancos y sistemas de justicia. A lo largo de este estudio, se ha demostrado que la IA no solo amplifica las capacidades de grupos delictivos, sino que permite la creación de nuevos entornos criminales —más autónomos, más difíciles de rastrear, más resistentes al control institucional. Las organizaciones criminales han comenzado a integrar modelos generativos, algoritmos de segmentación, sistemas de voz sintética y simulaciones judiciales para extorsionar, defraudar, reclutar, vigilar y desestabilizar. Frente a esta realidad, el tiempo de las advertencias ya ha pasado: lo que se necesita ahora es una agenda de acción.



A partir del trabajo de campo, el análisis de casos y las entrevistas estratégicas con autoridades de nueve países, se presenta una hoja de ruta que contiene dieciséis recomendaciones organizadas en cuatro ejes complementarios:

I. ACTUALIZACIÓN NORMATIVA: TIPIFICAR, REGULAR, ADAPTAR

Uno de los déficits más persistentes identificados es la falta de adecuación normativa frente al uso criminal de inteligencia artificial. Aunque algunos países han avanzado en leyes sobre ciberseguridad o delitos informáticos —como Chile con su Ley 21.459 o El Salvador con su ley de 2016—, la mayoría carece de figuras penales específicas que reconozcan los riesgos asociados al uso malicioso de modelos generativos, algoritmos de manipulación o evidencia sintética. Este eje propone reformas urgentes que no solo respondan al rezago normativo, sino que preparen a los sistemas de justicia para una transformación irreversible en la naturaleza del delito.

1. TIPIFICAR PENALMENTE LOS DELITOS ALGORÍTMICOS EMERGENTES

La legislación regional requiere una actualización profunda que incluya al menos cinco figuras penales emergentes:

- Clonación de voz con fines de extorsión, ya documentada en Ecuador, México y Colombia.
- Generación y distribución de contenido sintético (deepfakes) con fines de daño reputacional, extorsión o chantaje, como ocurre en El Salvador, Chile y Francia.
- Automatización del fraude mediante sistemas inteligentes, especialmente en esquemas tipo Montadeudas o phishing adaptativo, presentes en México, Brasil y Colombia.
- Manipulación algorítmica de emociones y percepciones, relevante en campañas de desinformación como las reportadas en Ecuador, Chile y Francia.
- Uso instrumental de IA en delitos de trata, vigilancia criminal o segmentación de víctimas.

Estas figuras deben ser tipificadas con autonomía suficiente para permitir su persecución penal, y acompañadas de agravantes cuando vulneren derechos fundamentales, involucren a grupos en situación de vulnerabilidad o afecten servicios estratégicos como procesos electorales, infraestructura crítica o seguridad nacional.

2. ESTABLECER MARCOS DE COOPERACIÓN OPERATIVA CON PLATAFORMAS DIGITALES

La persecución del crimen algorítmico depende de la capacidad estatal para interactuar con las infraestructuras privadas donde se aloja y distribuye. Se recomienda establecer convenios técnico-jurídicos con empresas como Meta, Telegram, TikTok, Google o Cloudflare, que incluyan:

- Protocolos de solicitud y preservación de datos ante evidencia sintética.
- Mecanismos de respuesta rápida frente a contenido automatizado dañino.
- Acompañamiento técnico en investigaciones penales complejas.

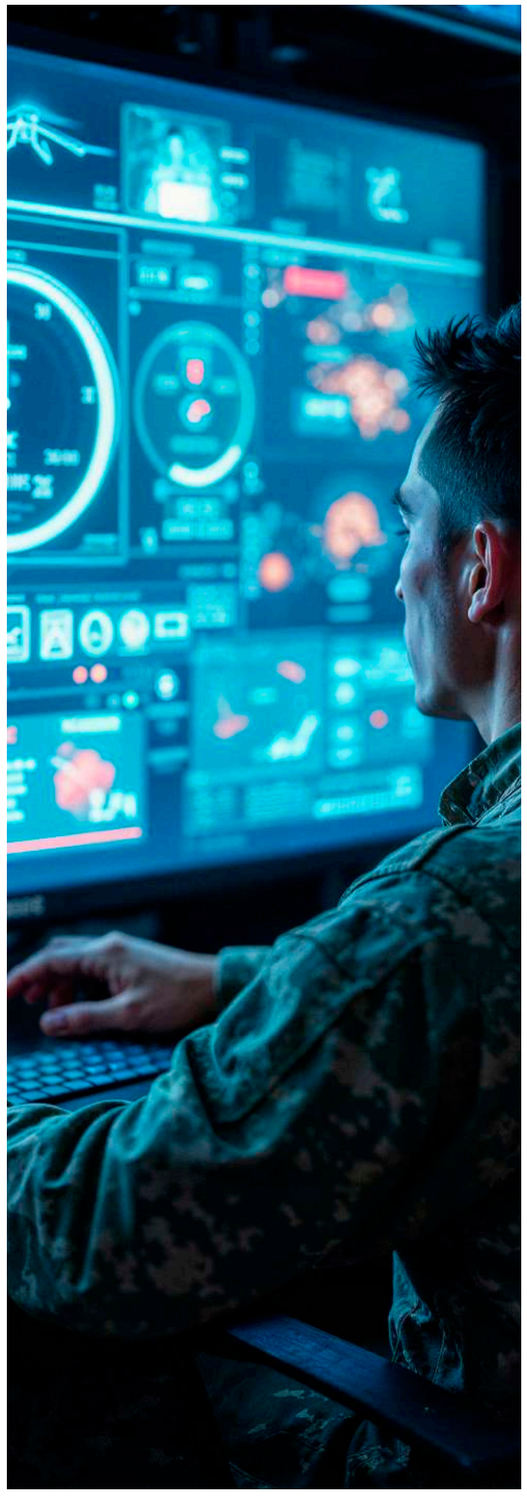
Se subraya la necesidad de formalizar estos acuerdos como condición mínima para el rastreo y atribución de contenido delictivo impulsado por IA.

3. ADAPTAR LAS GUÍAS TIPO SIRIUS AL CONTEXTO LATINOAMERICANO

Inspirados en el modelo europeo, países como Colombia y Francia propusieron la construcción de una guía SIRIUS-LAC que articule buenas prácticas para el manejo de evidencia digital generada por IA. Esta guía regional debería incluir:

- Criterios para la solicitud de información a plataformas tecnológicas.
- Estándares mínimos para la autenticación de evidencia sintética.
- Protocolos armonizados entre fiscales, jueces y peritos.

Este instrumento facilitaría el trabajo técnico-jurídico de operadores de justicia en la región y reduciría la asimetría institucional frente a actores transnacionales del crimen algorítmico.



II. FORTALECIMIENTO INSTITUCIONAL: CAPACIDADES PARA INVESTIGAR, ANALIZAR Y JUZGAR

Aunque algunas instituciones cuentan con unidades cibernéticas o expertos técnicos, el fenómeno del crimen algorítmico exige un salto cualitativo. No basta con tener policías informáticos: se requiere construir capacidades especializadas y transversales, integrando pericia técnica, razonamiento jurídico y visión estratégica. Este eje propone reforzar el músculo técnicoinstitucional con herramientas, protocolos, personal capacitado y estructuras funcionales.

4. CREAR UNIDADES ESPECIALIZADAS EN CRIMEN ALGORÍTMICO Y PLATAFORMAS DELICTIVAS AUTÓNOMAS

Las estructuras convencionales de cibercrimen son insuficientes para enfrentar las nuevas arquitecturas delictivas basadas en IA. Se propone crear unidades interinstitucionales con capacidades técnico-jurídicas avanzadas y enfoque multidisciplinario. Estas unidades deben integrar:

- Peritos digitales, fiscales, policías cibernéticos, ingenieros de datos y analistas de inteligencia.
- Facultades para investigar algoritmos generativos, segmentación automatizada de víctimas, plataformas Crime-as-a-Service y automatización del fraude.
- Mecanismos de coordinación con CERTs, unidades contra delitos financieros y cuerpos antiterroristas. Brasil, México y Perú destacan la urgencia de conformar estas células híbridas ante la convergencia de delitos financieros, cibernéticos y organizados.

5. DISEÑAR PROTOCOLOS FORENSES APLICABLES A EVIDENCIA GENERADA POR IA

El análisis de evidencia sintética requiere procedimientos específicos que garanticen su integridad, autenticidad y admisibilidad judicial. Se recomienda desarrollar protocolos nacionales que incluyan:

- Estándares de hash, autenticación de metadatos y consistencia semántica en textos, imágenes y audios generados por IA.
- Criterios de trazabilidad algorítmica y cadena de custodia para evidencias volátiles.
- Procedimientos probatorios diferenciados según el tipo de modelo generativo implicado.
- La ausencia de protocolos forenses obstaculiza la persecución penal de delitos con IA.

6. DOTAR A LAS INSTITUCIONES DE HERRAMIENTAS DE DETECCIÓN Y VERIFICACIÓN DE IA MALICIOSA

La respuesta institucional al crimen algorítmico requiere equipamiento técnico especializado. Se recomienda adquirir herramientas de detección con capacidades de:

- Identificación de deepfakes, bots generativos, clonación de voz y patrones textuales sintéticos.
- Verificación cruzada de imágenes, audios y videos con soluciones open-source adaptadas localmente.
- Atribución de contenido a modelos generativos específicos mediante análisis forense algorítmico.
- Inversión en soluciones tecnológicas de detección y atribución.

7, FORTALECER LA COORDINACIÓN TÉCNICO-JURÍDICA ENTRE FISCALÍAS Y CUERPOS POLICIALES

La investigación de delitos cometidos mediante IA depende de esquemas colaborativos claros entre operadores de justicia. Se propone:

- Revisar los protocolos de coinvestigación, solicitud de datos, recolección y preservación de evidencia algorítmica.
- Establecer lenguajes comunes entre técnicos y juristas para facilitar la colaboración diaria.
- Implementar marcos de interoperabilidad entre fiscalías, policías y unidades de análisis criminal.

8. CONSOLIDAR CAPACIDADES PERICIALES EN EVIDENCIA GENERADA POR IA

El uso de IA para fines delictivos exige nuevas competencias periciales en los sistemas de justicia. Se recomienda:

- Equipar laboratorios forenses y fiscalías con herramientas de autenticación y análisis de contenido sintético.
- Desarrollar metodologías para auditar modelos generativos, verificar contenido clonado y validar manipulación automatizada.
- Incluir criterios de verificación algorítmica y estandarización probatoria en los informes periciales.
- Ncesidad urgente de desarrollar capacidades forenses específicas en IA.

9. IMPULSAR LA FORMACIÓN JUDICIAL Y FISCAL EN EVIDENCIA SINTÉTICA

Los delitos con IA presentan desafíos inéditos para el análisis probatorio. Se propone establecer programas de formación regional para jueces, fiscales y defensores públicos que incluyan:

- Módulos sobre autenticidad, trazabilidad, cadena de custodia y estándares de admisibilidad de evidencia generada por IA.
- Simulaciones de análisis forense de contenido sintético y dictámenes técnico-jurídicos.
- Cooperación con instituciones como CEJA, IberRed y redes judiciales europeas para garantizar un enfoque comparado.
- Necesidad de actualizar las competencias judiciales ante la transformación del escenario probatorio.

III. COOPERACIÓN REGIONAL E INTERINSTITUCIONAL: ACTUAR EN RED, COMPARTIR CAPACIDADES

Frente a un fenómeno transnacional, descentralizado y tecnológicamente complejo, ninguna institución ni país puede enfrentar solo al crimen algorítmico. Las entrevistas revelaron un consenso: la cooperación —jurídica, operativa, técnica y política— es indispensable. Este eje propone mecanismos para facilitar el flujo de información, la interoperabilidad institucional y la creación de comunidades regionales de práctica.

10. CONSTRUIR UNA AGENDA REGIONAL DE GOBERNANZA CRIMINAL DE LA IA

Frente a un fenómeno transnacional y acelerado como el crimen algorítmico, se requiere una arquitectura de gobernanza compartida. Se recomienda:

- Establecer una agenda regional que articule alertas, estándares de evidencia, protocolos conjuntos y tipologías emergentes.
- Impulsar este esfuerzo desde mecanismos como AIAMP, IberRed, Ameripol y EL PACCTO.
- Incluir criterios éticos, interoperabilidad procesal y cooperación técnico-jurídica con plataformas digitales. Francia ofreció su experiencia como referente europeo en la construcción de sistemas compartidos de gobernanza sobre evidencia digital y algoritmos.

11. CREAR UNA BASE DE DATOS REGIONAL SOBRE INCIDENTES CRIMINALES CON IA

Para anticipar patrones y construir alertas tempranas, se propone establecer una base de datos regional que contenga:

- Casos verificados de uso delictivo de IA, clasificados por modus operandi, tecnología empleada y actores involucrados.
- Mecanismos de reporte institucional y análisis automatizado de tendencias criminales con modelos predictivos.
- Interfaz segura y accesible para operadores judiciales y cuerpos técnicos. Algunos países solicitaron una herramienta común para detectar, documentar y responder a las mutaciones del crimen algorítmico.

12. FORTALECER LA COORDINACIÓN MULTI-JURISDICCIONAL EN PAÍSES FEDERALES.

En países con estructuras federales como México y Brasil, es urgente articular respuestas coherentes entre niveles de gobierno. Se recomienda:

- Crear mesas técnicas permanentes que integren fiscales, policías, autoridades tecnológicas y actores legislativos.
- Emitir lineamientos operativos conjuntos, distribuir competencias y armonizar protocolos de preservación y atribución.
- Coordinar con fiscalías especializadas en delitos digitales, trata, crimen organizado y delitos financieros. Las brechas de interoperabilidad entre instancias locales y federales fueron señaladas como factor crítico por operadores en Brasil y México.

13. DESIGNAR PUNTOS DE CONTACTO NACIONALES ESPECIALIZADOS EN IA CRIMINAL (SPOC-IA)

Se recomienda que cada país designe una unidad de contacto que canalice la respuesta institucional frente a incidentes de IA maliciosa, con funciones como:

- Interlocución con plataformas tecnológicas y redes policiales internacionales (INTERPOL, IberRed, etc.).
- Coordinación de preservación de datos, solicitudes judiciales transfronterizas y alertas técnicas.
- Articulación con los CSIRT y las unidades nacionales de ciberseguridad. Perú, Brasil y Colombia destacaron la necesidad de contar con SPOCs para centralizar información técnica y legal.

14. ESTANDARIZAR Y HACER INTEROPERABLES LAS BASES DE DATOS SOBRE DELITOS CON IA

La región necesita construir sistemas compatibles de información criminal en IA. Se propone:

- Establecer metadatos estandarizados sobre incidentes algorítmicos.
- Conectar las bases nacionales con una plataforma regional de análisis predictivo y visualización de patrones.
- Desarrollar interfaces de acceso seguro para policías, fiscales y unidades técnicas. México, Ecuador y Colombia advirtieron que la dispersión de datos impide entender la escala y evolución del crimen algorítmico en la región.

IV. ENFOQUES EN DERECHOS HUMANOS, GÉNERO Y PROTECCIÓN DE VÍCTIMAS

El uso criminal de la IA no solo plantea desafíos tecnológicos o legales, sino también problemas éticos, sociales y políticos. Los Montadeudas automatizados, las campañas de acoso digital o los deepfakes sexuales afectan de manera desproporcionada a mujeres, niños, comunidades rurales o personas con baja alfabetización digital. Este eje propone medidas de prevención, protección y formación con perspectiva de derechos.

15. DISEÑAR CAMPAÑAS PÚBLICAS DE ALFABETIZACIÓN DIGITAL SOBRE IA CRIMINAL

La protección ciudadana frente al crimen algorítmico empieza por el conocimiento. Se recomienda:

- Desarrollar campañas accesibles y multilingües dirigidas a comunidades vulnerables: mujeres, personas mayores, adolescentes y pueblos indígenas.
- Incluir ejemplos reales de fraudes con voz clonada, montadeudas, videos manipulados o mensajes automatizados.
- Enseñar señales de alerta, rutas de denuncia y recomendaciones prácticas. Muchas víctimas no reportan los ataques por desconocimiento o estigmatización.

16. ESTABLECER SALVAGUARDAS DE DERECHOS HUMANOS EN SISTEMAS AUTOMATIZADOS DE VIGILANCIA

El uso de IA para prevenir delitos debe observar principios jurídicos fundamentales. Se recomienda:

- Someter a auditorías independientes los sistemas de análisis predictivo, reconocimiento facial o detección algorítmica.
- Incluir mecanismos de revisión humana obligatoria y principios de proporcionalidad, legalidad y no discriminación.
- Integrar estos controles en el diseño de algoritmos utilizados por las autoridades estatales.

CONCLUSIONES

El de inteligencia artificial por organizaciones criminales de alto riesgo ya no es una posibilidad hipotética ni un fenómeno en fase embrionaria. Es una realidad operativa que está reconfigurando los métodos, estructuras y alcances del delito organizado. El presente estudio confirma que las redes criminales —sean cárteles, mafias carcelarias, grupos paramilitares, proxies estatales o colectivos cibernéticos— están incorporando tecnologías algorítmicas no sólo para optimizar sus actividades, sino para expandir sus capacidades de control social, manipulación simbólica, segmentación de víctimas y evasión de la acción penal.

Los hallazgos revelan una multiplicidad de dinámicas criminales donde la IA se integra como catalizador delictivo: clonación de voz para fraudes telefónicos; uso de deepfakes con fines de extorsión, daño reputacional o chantaje político; estafas automatizadas mediante bots conversacionales; manipulación emocional mediante sistemas de segmentación de públicos; suplantaciones digitales de plataformas legítimas; vigilancia criminal por reconocimiento facial y minería de datos biométricos; así como generación de contenido sintético para campañas de desinformación o captación de víctimas.

En todos los casos estudiados, la inteligencia artificial redefine escala y velocidad operativa del crimen. Permite hacer más con menos: más daño, más víctimas, más control, más impunidad, con menos exposición humana, menos recursos y menos trazabilidad. Esta lógica no sólo fortalece a las organizaciones delictivas, sino que transforma el entorno mismo de la justicia penal: los fiscales, jueces, policías, defensores públicos y peritos enfrentan escenarios para los cuales no han sido entrenados ni institucional ni jurídicamente.

Una conclusión transversal es que las capacidades institucionales siguen rezagadas frente al ritmo de adopción tecnológica por parte de redes criminales. En la mayoría de los estudio de caso y en los países analizados, las legislaciones penales siguen sin tipificar figuras como la automatización del fraude, la manipulación algorítmica o la evidencia sintética. Las fiscalías carecen de herramientas para verificar contenido generado por IA, las policías no tienen acceso oportuno a plataformas tecnológicas y los jueces enfrentan desafíos inéditos en la admisión, autenticación y valoración de pruebas digitales. A esto se suma una marcada desigualdad territorial: mientras algunas capitales disponen de unidades forenses especializadas, las regiones periféricas no cuentan siquiera con conectividad o personal técnico capacitado.

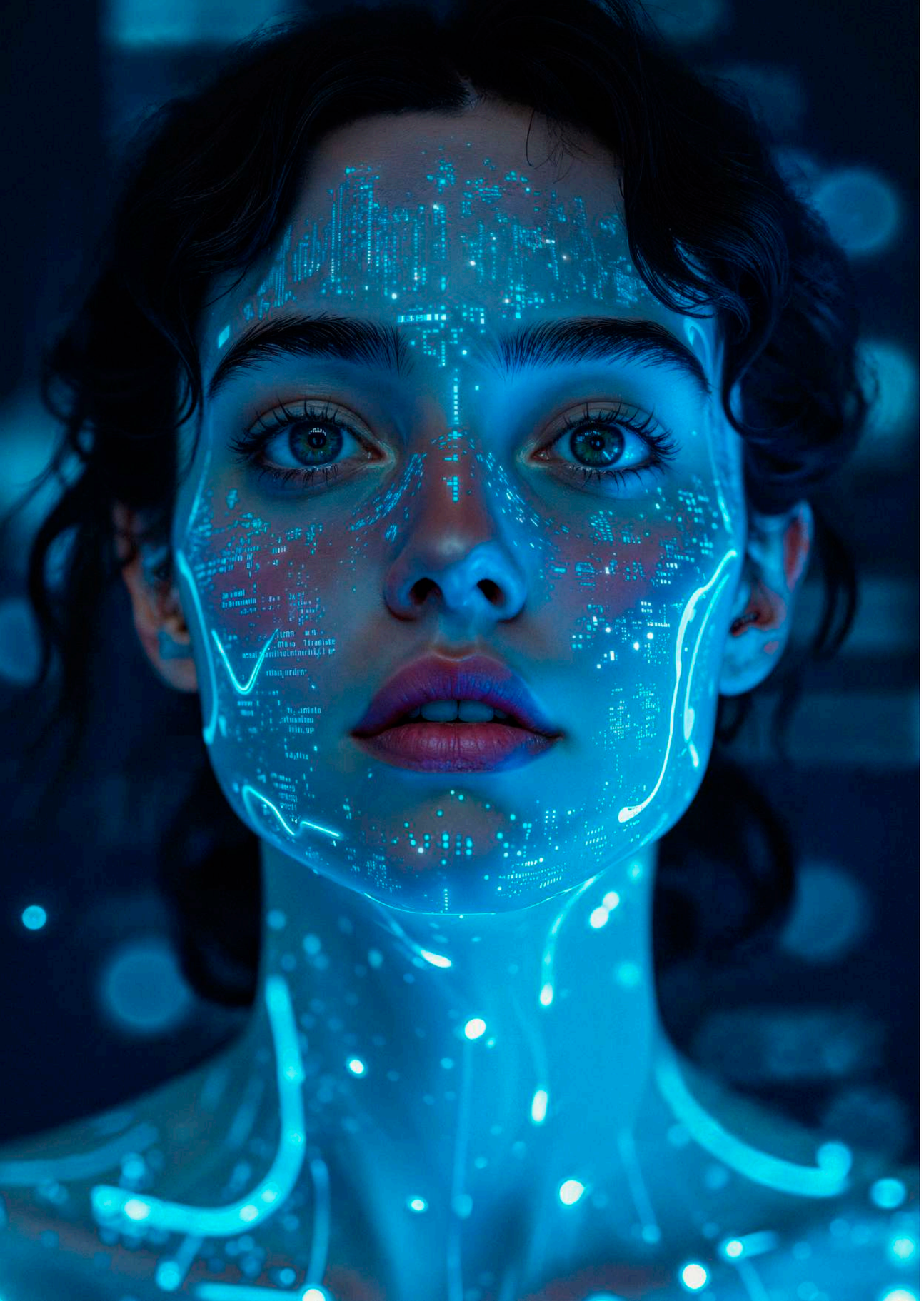
Otro hallazgo clave es que la criminalidad algorítmica no se limita a actores individuales ni a grupos puramente privados. El estudio documentó operaciones impulsadas por proxies estatales, actores paraestatales y ecosistemas híbridos donde confluyen intereses geopolíticos, plataformas digitales y operadores criminales. En estos casos, la IA no se usa solo para cometer delitos, sino para deslegitimar adversarios, erosionar la confianza pública, generar caos informativo o manipular procesos democráticos. Se trata de una confrontación que no se libra por el control de territorios físicos, sino por la arquitectura cognitiva de la realidad social.

Frente a esta transformación, el estudio propone una hoja de ruta integral organizada en cuatro ejes: actualización normativa, fortalecimiento institucional, cooperación regional e interinstitucional, y protección de derechos humanos y víctimas. Entre las medidas prioritarias destacan: tipificar penalmente el uso criminal de IA, establecer protocolos forenses para evidencia sintética, crear unidades especializadas en crimen algorítmico, adaptar las guías SIRIUS al contexto latinoamericano, invertir en herramientas de detección y verificación algorítmica, establecer puntos de contacto técnico-jurídicos nacionales, armonizar bases de datos interoperables y garantizar auditorías éticas en los sistemas de vigilancia automatizada.

América Latina y Europa no parten de cero. Algunos países han dado pasos importantes, como Chile con su nueva Ley de Ciberseguridad (2024), Brasil con su ecosistema de ciberinteligencia policial, o Francia como referente comparado en estándares forenses y cooperación judicial transnacional. Sin embargo, la región aún carece de una estrategia compartida, multilateral y funcional frente al crimen algorítmico. La cooperación técnica, el intercambio de buenas prácticas, la estandarización probatoria y la formación judicial especializada deben ser concebidas como condiciones de soberanía jurídica frente a una criminalidad que ya no necesita armas ni ejércitos, sino código, infraestructura y anonimato.

Finalmente, el estudio alerta sobre un riesgo estructural: si los Estados no comprenden, enfrentan y regulan de forma estratégica la convergencia entre crimen e inteligencia artificial, quedarán subordinados a un orden delictivo que no reconoce fronteras, ni valores, ni contrapesos. El crimen algorítmico no solo produce víctimas; produce exclusión digital, impunidad programada, incertidumbre jurídica y deslegitimación institucional. Lo que está en juego no es únicamente la seguridad, sino la integridad misma de nuestros sistemas democráticos de justicia.

El futuro del combate al delito en América Latina no puede construirse con paradigmas del pasado. Se requiere un nuevo lenguaje penal, una arquitectura técnica interoperable, y una cooperación que trascienda ministerios y fronteras. Este estudio no ofrece certezas absolutas, pero sí una brújula operativa. Porque cuando el crimen aprende más rápido que la justicia, no basta con reaccionar: hay que anticiparse.



AGRADECIMIENTOS

El autor desea expresar su más profundo agradecimiento a las instituciones y personas que hicieron posible este estudio a través de su generosa disposición, tiempo y conocimientos especializados.

En Brasil, se reconoce la colaboración de Vanessa Goncalvez Leite de Souza de la Policía Federal, junto a la Perita Criminal Federal Maria Isabel Vasconcelos Lima. También agradezco la participación de Ricardo Magno de Texeira del Tribunal de Justiça do Distrito Federal e dos Territórios.

En Chile, agradezco a Juan Pablo Glasinovic Vernon de la Fiscalía Nacional, así como a Claudio Ramírez, Marcela Toledo, Francisco Andaur, Tania Gajardo y al propio Juan Pablo Glasinovic por su colaboración desde distintas unidades especializadas. Asimismo, extiendo un especial reconocimiento a Claudia Moyano Navarrete del Ministerio del Interior, y a los equipos operativos de la Policía de Investigaciones (PDI) —Subprefecta Jazmín Cárdenas, Comisario Jonathan Castillo, Inspector Esteban Donoso, Subinspector Ignacio Cárcamo y Comisario Danic Maldonado—, así como al Capitán Juan Pablo Lastra, Capitán Felipe Cáceres y José Garrido de Carabineros de Chile.

En Colombia, se reconoce la participación activa de Diana Catalina Calderón Millán del Ministerio de Defensa, así como del CT Óscar Iván Mendoza García y del ST Edward Gonzalo López Mejía.

En Ecuador, agradezco profundamente a Luis Fabián Armijos Samaniego del Ministerio del Interior y al amplio equipo interinstitucional que participó activamente: Juan Ávila, Ariana Zambrano, Angelita Severino, Alex Carcelén, Erick Banegas, Shirley Galarza, Jayder Chala, Edwar Chala, José Vilañez, Jonathan Flores, Fernando Mullo, Diego Taipe, Carlos Vélez, Darwin Toro y Javier López.

En El Salvador, el estudio se benefició del respaldo institucional de Romeo Vargas del Ministerio de Justicia y Seguridad Pública, así como de Jaime Perla Flores, Inspector Jefe Gerardo Bonilla Solano y sus respectivos equipos de la Policía Nacional Civil.

En Francia, el autor agradece la disposición, análisis comparado y aportes técnicos brindados por Yann Loubry y Simon Paul del Ministère de la Justice, quienes facilitaron un puente clave entre experiencias europeas y necesidades regionales latinoamericanas.

En México, se agradece la valiosa interlocución con Israel Agüero (SSPC), Jesús Hernández (Policía Cibernética), Miguel Báez (CNI) y Patricia Chávez Obregón, por sus aportaciones sustantivas.

En Perú, se agradece el apoyo de Silvia Nayda De la Cruz Quintana del Ministerio del Interior, y del General PNP José Antonio Zavala Chumbiauca por su participación y análisis estratégico.

Finalmente, un agradecimiento especial a EL PACCTO 2.0, por su apoyo técnico, institucional y logístico a lo largo de esta investigación. En particular, a Marc Reina Tortosa y a Emily Breyne, cuyo acompañamiento riguroso y constante resultó indispensable para el desarrollo metodológico y el enfoque estratégico del estudio.

Este documento es el resultado de un esfuerzo colectivo, guiado por el compromiso compartido de fortalecer las capacidades institucionales frente al crimen con uso de inteligencia artificial.

BIBLIOGRAFÍA

Acertpix. (2025, February 18). KK Park: The online fraud factory involved in employee exploitation.

ADN40. (2024). Predatory loan apps in Mexico 2024: Complete list and how to avoid scams.

Agencia Boliviana de Información. (2025, February 10). Criminal organization cloned Minister Véliz's voice with AI, defrauded 19 people by selling positions and obtained over Bs 5 million.

Aguilar Antonio, J.M. (2024). Ransomware gangs and hacktivists: Cyber threats to governments in Latin America. Florida International University, Jack D. Gordon Institute for Public Policy.

Ahmed, D. (2025, April 7). Xanthorox AI Surfaces on Dark Web as Full Spectrum Hacking Assistant. Hackread.

AIID (2025) Incident reports 690, 725, 727, 897, 901, 911, 912, 913, 918, 929, 937, 955, 958 y 1015: Reported darknet launch of Xanthorox AI introduces autonomous cyberattack platform.

Alvarado Flores, M.E. (2025, February 10). Criminal organization used artificial intelligence to simulate the voice of the Minister of Education and commit fraud. Visión 360.

Anggorojati, B., Perdana, A., Wijaya, D. (2024, July 24). FraudGPT and other malicious AIs are the new frontier of online threats. What can we do? The Conversation.

Atanasova, A., Reset Tech, Check First. (2025, July 1). A pro-Russia disinformation campaign is using free AI tools to fuel a content explosion. Wired.

Bangkok Post. (2025, March 2). Two men arrested for alleged B4m AI-aided scam against beauty queen.

Barman, D., Guo, Z., Conlan, O. (2024). The dark side of language models: Exploring the potential of LLMs in multimedia disinformation generation and dissemination. Machine Learning with Applications.

Barragán, C. (2023, July 11). Inside the world of Nigerian Yahoo boys. Longreads / The Atavist Magazine.

Bayer, J., Pineda, J., Li, Y. (2024, January 30). How Chinese mafia are running a scam factory in Myanmar. DW.

Béchard, D. E. (2025, May 7). *Xanthorox AI lets anyone become a cybercriminal*. Scientific American.

Bitdefender Enterprise. (2025, March 4). FunkSec: An AI-centric and affiliate-powered ransomware group.

Burton, J., Janjeva, A., Moseley, S., Alice. (2025). AI and serious online crime. Centre for Emerging Technology and Security (CETaS), The Alan Turing Institute.

C4ADS. (2025, March 27). Hot lines: Tracing movements to and from Myanmar's scam centers.

Caldwell, M., Andrews, J.T.A., Tanay, T., Griffin, L.D. (2020). AI-enabled future crime. Crime Science 9, 14.

Caulfield, J. (2024). The Yahoo-boys and the upsurge in sextortion – Part 1 & 2. Linkedin. Check Point Software. (2025, May). FunkSec ransomware – AI powered group.

Cheng, N. (2025, March 25). National police capture Thai ringleaders during Poipet scam raids. The Phnom Penh Post.

Chukwuma, O.K. (2024). Understanding the crime-grid of the Nigerian Yahoo boys. National Journal of Cyber Security Law 7(2).

Consejo Ciudadano para la Seguridad y Justicia CDMX. (2022). Montadeudas typology: Analysis and recommendations.

CybelAngel. (2023). The dark side of Gen AI: Uncensored large language models [white paper].

CybelAngel. (2025). Gen AI and the rise of uncensored LLMs on the dark web.

Cyber Florida at University of South Florida. (2025, January 29). FunkSec: A top ransomware group leveraging AI.

Dark Reading. (2024, November 5). Iranian APT targets IP cameras, extends attacks beyond Israel.

Deibert, R. (2023). Reset: Reclaiming the internet for civil society. House of Anansi Press.

Der Spiegel. (2025, February 12). German election campaign flooded with fake news and videos. Der Spiegel International.

Di Girolamo, M. (2025, March 27). Hot lines: Tracing movements to and from Myanmar's scam centers. C4ADS.

Dueñas, D. (2023, June 26). How to avoid predatory loan scams. Capital 21.

Durán San Juan, I. (2024, October 4). This is how cybercriminals use AI to scam people in Latin America: How you can protect yourself. Infobae.

El Deber. (2025, February 10). Criminal organization dismantled after using the voice of the Minister of Education to defraud.

Enterprise Security Tech. (2025, April 8). Russia's "Pravda" disinformation network is poisoning Western AI models.

Enterprise Security Tech. (2025, March 2). Microsoft names developers behind AI jailbreaking tools in legal crackdown on Storm-2139.

Europol. (2024). Decoding the EU's most threatening criminal networks. Publications Office of the European Union.

Europol. (2025). Child sexual exploitation. European Union Agency for Law Enforcement Cooperation.

Europol. (2025). EU SOCTA 2025: Strategic report on serious and organised crime in the European Union. Europol

FDD. (2024, October 24). America resilient in the face of aggressive foreign malign influence targeting the 2024 U.S. elections.

FireXCore. (2025, May 25). AI-driven ransomware FunkSec: The shocking fusion of hacktivism and cybercrime.

García, S. (2025, May 8). How criminal groups have adapted to the digital age. InSight Crime.

GITOC. (2023). Global organized crime index 2023. Global Initiative Against Transnational Organized Crime

GNET. (2024). AI-powered jihadist news broadcasts: A new trend in pro-IS propaganda production. Global Network on Extremism and Technology.

Griffin, M. (2025, April 26). Revolutionary autonomous cyberattack platform emerges on the dark web. Fanatical Futurist.

Head, J. (2025, February 15). Scams, casinos and skyscrapers: The luxurious ghost city that emerged in one of the world's poorest areas (and in the middle of a civil war). BBC News Mundo.

Infosecurity Magazine. (2024, November 6). US and Israel warn of Iranian threat actor's new tradecraft.

Iyer, P. (2024, January 18). Studying underground market for large language models, researchers find OpenAI models power malicious services. Tech Policy Press.

Johnson, D.B. (2025, February 27). Microsoft IDs developers behind alleged generative AI hacking-for-hire scheme. CyberScoop.

Kelley, D. (2025, April 7). Xanthorox AI – The next generation of malicious AI threats emerges. SlashNext.

Kiripost. (2025, March 26). Raids on Poipet scam centres find 63 Thais involved in online fraud.

Kykyo (2024). *Chinese criminal gangs drive rise in pig-butchering scams as victims suffer emotional, financial harm* Coinlive.

Lakshmanan, R. (2024). Inside Iran's cyber playbook: AI, fake hosting, and psychological warfare. The Hackers News.

López Ponce, J. (2025, January 27). How digital predatory loan scams operate in Mexico: UIF combats psychological extortion Black Mirror style. Milenio.

Lyngaas, S. (2021, agosto 9). Arbitration among cybercriminals: Inside the underground world of XSS, Exploit and REvil ransomware. CyberScoop.

Marsh, S. (2025, January 20). Russian disinformation targets German election campaign, says think tank. Reuters.

Martínez A. (2023, June 26). Debt app detainees avoid pretrial detention. Milenio:

Martínez, R. (2024, August 27). This is how the CJNG uses AI to commit fraud and extortion, according to InSight Crime. Infobae.

Martínez, R. (2024, May 8). These are the apps used by the Sinaloa Cartel and Los Chapitos to communicate without leaving a trace. Infobae.

Masada, S. (2025, February 27). Disrupting a global cybercrime network abusing generative AI. Microsoft On the Issues.

McCready, A., Mendelson, A. (2023, July 22). Myanmar: Chinese-run scam hubs reportedly continue running unabated with signs of human trafficking and forced labour. Business & Human Rights Resource Centre.

Menn, J. (2025, April 17). Russia seeds chatbots with lies. Any bad actor could game AI the same way. The Washington Post.

Microsoft. (2024a, October 23). As the U.S. election nears, Russia, Iran and China step up influence efforts. Microsoft On the Issues.

Microsoft. (2025, February 29). Microsoft disrupts Storm-2139 for LLMjacking and Azure AI exploitation.

Ministère de l'Europe et des Affaires étrangères. (2024, February 15). Foreign digital interference – Result of investigations into the Russian propaganda network Portal Kombat.

Ministerio de Educación de Bolivia. (2025, February 10). Criminal organization used artificial intelligence to clone the voice of the Minister of Education, Omar Véliz Ramos.

MITRE. (2025). ATLAS™: Adversarial Threat Landscape for Artificial-Intelligence Systems. MITRE Corporation.

Narim, K. (2025, February 24). Cambodian police raid scam centers in Poipet, discover over 200 foreigners. CamboJA News.

Nath, S. (2025, April 13). *This AI tool empowers cybercriminals with advanced capabilities—No jailbreaks needed*. The420.in.

NewsGuard. (2025). Russia's "Pravda" network poisons AI training data.

Newton, C. (2024, August 26). How AI is transforming organized crime in Latin America. InSight Crime.

Nicholls, C. (2025, February 28). Dozens arrested in crackdown on AI-generated child sexual abuse material. CNN.

Nilsson Julien, E. (2025, February 7). Fake TikTok videos show hundreds of thousands marching for AfD in Germany. Euronews.

Ojedokun, U.A., Ilori, A.A. (2021). Tools, techniques and underground networks of Yahoo-boys in Ibadan City, Nigeria. International Journal of Criminal Justice 3, 99–122.

Oloworekende, A. (2019, August 28). Yahoo Yahoo – Nigeria and cybercrime's global ecosystem. The Republic.

Orgaz, C.J. (2024, October 4). Artificial intelligence: 6 ways Latin American criminal groups use AI to commit crimes. BBC News Mundo.

Partnership on AI. (2022). Report on algorithmic risk assessment tools in the U.S. criminal justice system.

Penang Institute. (2023). Combating scam syndicates in Malaysia and Southeast Asia. Penang Institute Policy Brief.

PlasBit (Ziken Labs). (2024, July 7). What is KK Park Myanmar: Crypto scams and human trafficking.

Poireault, K. (2023). The dark side of generative AI: Five malicious LLMs found on the dark web. Infosecurity Europe.

Racoveanu, C. (2024). Artificial intelligence – a double-edged sword: Organized crime's AI vs law enforcement's AI. In Proceedings of the 18th International Conference on Business Excellence, 408–419. ASE Publishing.

Raksmey, H. (2025, February 24). Poipet scam compound raids net 230 foreigners, more rescued. The Phnom Penh Post.

Regan, H., Watson, I., Rebane, T., Olarn, K. (2025, April 2). Global scam industry evolving at unprecedented scale despite recent crackdown. CNN.

Rosiek, T. (2025, March 21). Data poisoning threatens AI's promise in government. FedTech Magazine.

Ruvnet. (2024). The emergence of malicious large language models (LLMs) and the next frontier of symbolic-AI integration. GitHub.

Schultz, J. (2024, junio 4). Cybercriminal abuse of large language models. Talos Intelligence. Cisco Talos.

Secretaría de Hacienda y Crédito Público. (2024). National risk assessment on money laundering and terrorist financing.

SlashNext. (2025). Xanthorox AI – The next-gen malicious AI.

SOCRadar. (2023, diciembre 4). *Under the spotlight: RAMP forum*. SOCRadar Threat Intelligence Blog.

SOCRadar. (2025, January 4). Dark web profile: FunkSec. SOCRadar Cyber Intelligence Inc.

Speckhard, A., Thakkar, M. (2024, July 15). ISIS supporters harness the power of AI to ramp up propaganda on Facebook, X and TikTok. Homeland Security Today.

THAI.NEWS. (2025, February 3). Charlotte Austin's 4 million baht loss: Inside the Poipet call scam bust in 2025.

Tharayil, R. (2025, February 28). Microsoft expands legal action against AI abuse network Storm-2139. Tech Monitor.

The Nation Thailand. (2025, March 3). 119 Thais from Poipet: Victims or accomplices in a call centre scam?

The Record. (2024, October 31). FBI: Iranian cyber group targeted Summer Olympics with attack on French display provider.

Times of India. (2024, abril 3). News Harvest: How Islamic State is using AI anchors to boost propaganda.

TRM Labs. (2024, July 26). Authorities unravel the Sinaloa Cartel's connection to Chinese money launderers. TRM Blog.

TRM Labs. (2025). The rise of AI-enabled crime: Exploring the evolution, risks, and responses to AI-powered criminal enterprises.

UNICRI. (2021). Algorithms and terrorism: The malicious use of artificial intelligence for terrorist purposes.

UNODC. (2022). Digest of cyber organized crime: Second edition. United Nations.

Varese, F. (2010). What is organised crime? In F. Varese (Ed.), Organized crime: Critical concepts in criminology (Vol. 1, pp. 11–33). Routledge.

Vectra AI. (2025, May). Is your organization safe from FunkSec?

Vongthongsri, K. (2025, March 15). How to jailbreak LLMs one step at a time: Top techniques and strategies. Confident AI.

Wall, D.S. (2015). Dis-organised crime: Towards a distributed model of the organization of cybercrime. The European Review of Organised Crime 2, 71–90.

Whelan, C., Bright, D., Martin, J. (2024). Reconceptualising organised (cyber)crime: The case of ransomware. Journal of Criminology 57, 45–61.

Willsher, K., O'Carroll, L. (2024, February 12). French security experts identify Moscow-based disinformation network. The Guardian.

Ziken Labs. (2024, julio 7). What Is KK Park Myanmar: Crypto Scams and Human Trafficking. PlasBit.

EL PACCTO 2.0

EU-LAC Partnership on justice and security