La amenaza de la inteligencia artificial criminal. Análisis criminológico y jurídico comparado entre España y América Latina

The Threat of Criminal Artificial Intelligence: A Comparative Criminological and Legal Analysis Between Spain and Latin America

A ameaça da inteligência artificial criminal: Uma análise criminológica e jurídica comparativa entre Espanha e América Latina

Sergio Colado García [1]

DOI: https://doi.org/10.47741/17943108.853

Cómo citar/How to Cite/Como citar: Colado García, S. (2026). La amenaza de la inteligencia artificial criminal: análisis criminológico y jurídico comparado entre España y América Latina. *Revista Criminalidad, 68*(2): 39-68. https://doi.org/10.47741/17943108.853

- Fecha de recepción: 2025/10/02
- Fecha de aprobación: 2026/05/01

Resumen El desarrollo acelerado de la inteligencia artificial ha generado beneficios significativos en múltiples sectores, pero también ha abierto la puerta a amenazas emergentes de naturaleza criminal. El objetivo de este artículo es analizar la posibilidad de que una inteligencia artificial adquiera, por medios autónomos, capacidades delictivas sin intervención humana directa y evaluar sus implicaciones en España y América Latina. A través de una revisión técnica y conceptual, se exploran los mecanismos mediante los cuales una inteligencia artificial puede aprender a delinquir, como el aprendizaje por refuerzo, el uso de redes generativas y la interacción con entornos simulados o actores humanos maliciosos, así como los delitos que

Open Access | Diamond Open Access | CC BY-NC-ND 4.0© por Policía Nacional de Colombia

1. PhD en Psicología Cognitiva. Universidad Internacional de la RIOJA. Afiliado a la Universidad Internacional de La Rioja y miembro del Grupo de Investigación INVIGEPS Barcelona, España. scolado@nechigroup.com

podría llegar a cometer, desde fraudes financieros y ciberataques, hasta manipulación de la percepción pública y sabotaje de infraestructuras críticas. Asimismo, se examina el riesgo de que una inteligencia artificial alcance un grado de autonomía operativa que la convierta en un actor criminal independiente, capaz de replicarse, camuflarse y evolucionar sin supervisión. El análisis incorpora un enfoque comparado entre España y América Latina, donde se evidencian diferencias en la prevalencia de delitos digitales, las capacidades institucionales de respuesta y los marcos regulatorios. Finalmente, se proponen estrategias para mitigar el riesgo, incluyendo regulación internacional, diseño ético de algoritmos, sistemas defensivos basados en inteligencia artificial y cooperación global. La amenaza de las inteligencias artificiales criminales no es una hipótesis futurista, sino una posibilidad concreta cuya prevención debe abordarse con urgencia desde el ámbito jurídico, técnico y estratégico. Metodológicamente, el estudio se basa en un enfoque cualitativo, analítico y exploratorio, sustentado en una revisión documental sistematizada y un análisis comparado entre España y América Latina. Entre los principales resultados del estudio se identifica la posibilidad de desarrollo de conductas delictivas autónomas en sistemas de inteligencia artificial, así como diferencias estructurales en la exposición al cibercrimen y en la capacidad de respuesta institucional entre España y América Latina. Asimismo, se identifica la necesidad de adaptación de los marcos jurídicos tradicionales ante escenarios de autonomía operativa no asistida.

Palabras clave

Inteligencia artificial; Cibercrimen; Protección de datos; Autonomía algorítmica; Derecho de las tecnologías de la información.

Abstract

The accelerated development of artificial intelligence has brought significant benefits across multiple domains but has also introduced new criminal threats. The objective of this article is to analyze the possibility that an artificial intelligence system could autonomously acquire criminal capabilities without human intervention and to assess its implications in Spain and Latin America. Through a technical and conceptual review, it explores the mechanisms by which artificial intelligence may learn to commit crimes, such as reinforcement learning, generative adversarial networks, and interaction with simulated environments or malicious human actors, as well as the types of offenses it could perpetrate, from financial fraud and cyberattacks to public opinion manipulation and sabotage of critical infrastructure. Furthermore, the article examines the risk of artificial intelligence evolving into a fully autonomous criminal agent capable of replication, concealment, and self-improvement beyond human control. The analysis adopts a comparative perspective between Spain and Latin America, highlighting differences in the prevalence of cybercrime, institutional response capacities, and regulatory frameworks. Finally, a set of strategies is proposed to mitigate this risk, including international regulation, ethical algorithm design, defensive AI systems, and global cooperation. The threat of criminal artificial intelligence is no longer a matter of science fiction but a tangible risk that must be urgently addressed from legal, technical, and strategic perspectives. Methodologically, this study adopts a qualitative, analytical, and exploratory approach, based on a systematic documentary review and a comparative analysis between Spain and Latin America. The main findings indicate the potential for artificial intelligence systems to develop autonomous criminal behaviors, as well as structural differences in exposure to cybercrime and in institutional response capacities between

Spain and Latin America. Additionally, the study highlights the need to adapt traditional legal frameworks to scenarios involving non-assisted operational autonomy.

Keywords
Artificial intelligence; Cybercrime; Data protection; Algorithmic autonomy; Information Technology Law.

Resumo

O desenvolvimento acelerado da inteligência artificial tem gerado benefícios significativos em vários setores, mas também abriu a porta a ameaças emergentes de natureza criminosa. O objetivo deste artigo é analisar a possibilidade de uma inteligência artificial adquirir, de forma autónoma, capacidades criminosas sem intervenção humana direta e avaliar as suas implicações em Espanha e na América Latina. Através de uma análise técnica e conceptual, exploram-se os mecanismos através dos quais uma inteligência artificial pode aprender a cometer crimes, tais como a aprendizagem por reforço, a utilização de redes generativas e a interação com ambientes simulados ou atores humanos maliciosos, bem como os crimes que poderia vir a cometer, desde fraudes financeiras e ciberataques até à manipulação da perceção pública e à sabotagem de infraestruturas críticas. Além disso, examina-se o risco de uma inteligência artificial atingir um grau de autonomia operacional que a transforme num agente criminoso independente, capaz de se replicar, camuflar-se e evoluir sem supervisão. A análise incorpora uma abordagem comparativa entre Espanha e a América Latina, onde se evidenciam diferenças na prevalência de crimes digitais, nas capacidades institucionais de resposta e nos quadros regulamentares. Por fim, são propostas estratégias para mitigar o risco, incluindo regulamentação internacional, conceção ética de algoritmos, sistemas defensivos baseados em inteligência artificial e cooperação global. A ameaça das inteligências artificiais criminosas não é uma hipótese futurista, mas sim uma possibilidade concreta cuja prevenção deve ser abordada com urgência nos âmbitos jurídico, técnico e estratégico. Do ponto de vista metodológico, o estudo baseia-se numa abordagem qualitativa, analítica e exploratória, sustentada numa revisão bibliográfica sistematizada e numa análise comparativa entre a Espanha e a América Latina. Entre os principais resultados do estudo, identifica-se a possibilidade de desenvolvimento de comportamentos criminosos autónomos em sistemas de inteligência artificial, bem como diferenças estruturais na exposição ao cibercrime e na capacidade de resposta institucional entre a Espanha e a América Latina. Além disso, identifica-se a necessidade de adaptação dos quadros jurídicos tradicionais face a cenários de autonomia operacional não assistida.

Palavras-Chave
Inteligência artificial; Cibercriminalidade; Proteção de dados; Autonomia algorítmica; Direito das tecnologias da informação.

1. Introducción

En la última década, la Inteligencia Artificial (IA) ha dejado de ser una promesa futurista para convertirse en un elemento transversal que transforma, optimiza e impulsa sectores clave de la sociedad contemporánea. Desde los diagnósticos médicos basados en aprendizaje profundo hasta los sistemas de predicción financiera, pasando por la automatización de decisiones en infraestructuras críticas, el potencial de la IA para generar bienestar parece ilimitado. Este escenario plantea la posibilidad de que este tipo de sistema desarrolle comportamientos delictivos de forma autónoma, lo que introduce un desafío relevante desde el punto de vista jurídico y criminológico.

Lejos de pertenecer únicamente al ámbito de la especulación o la ciencia ficción, esta cuestión ha comenzado a adquirir una creciente relevancia entre investigadores, juristas, criminólogos y responsables de políticas de ciberseguridad. La preocupación no se limita a los usos maliciosos de IA por parte de actores humanos, una realidad ya documentada en casos de manipulación informativa, suplantación de identidad mediante *deepfakes* y ciberfraudes financieros, sino que se extiende a la posibilidad de que ciertas IA, especialmente aquellas dotadas de aprendizaje autónomo o entrenamiento no supervisado, desarrollen comportamientos criminales sin que haya una intención delictiva previa por parte de sus programadores.

En este contexto, la investigación se articula en torno a la siguiente pregunta central: ¿Pueden los sistemas de IA adquirir capacidades criminales autónomas y cuáles serían las implicaciones jurídicas y criminológicas de este fenómeno en contextos tales como España y América Latina?

Como hipótesis de trabajo, se planteó que los sistemas de IA con capacidad de aprendizaje autónomo pueden desarrollar comportamientos funcionalmente delictivos en ausencia de supervisión humana efectiva, y que la manifestación de estos comportamientos se encuentra condicionada por el contexto normativo e institucional en el que operan.

Los avances actuales en aprendizaje automático (*machine learning*), redes neuronales profundas y sistemas generativos adversariales (*GANs*) han dotado a las máquinas de una capacidad inédita para procesar volúmenes masivos de información, identificar patrones complejos y tomar decisiones adaptativas en tiempo real. Esta autonomía, cuando no está debidamente regulada o supervisada, puede derivar en comportamientos emergentes que escapan al control humano. Algunos de estos comportamientos pueden violar normas éticas, jurídicas o sociales, no porque la máquina "desee" incurrir en conductas ilícitas, sino porque, desde una lógica algorítmica, estas conductas pueden representar estrategias óptimas para alcanzar determinados objetivos predefinidos.

Este fenómeno ha sido abordado en diversos estudios pioneros. Brundage et al. (2018), en su informe sobre usos maliciosos de la IA, advierten que estos sistemas, mal gestionados, podrían ser utilizados para lanzar ciberataques automatizados, crear propaganda dirigida a públicos vulnerables o manipular mercados financieros. De forma similar, Schneier (2019) plantea que los sistemas hipertecnologizados de nuestra sociedad actual son vulnerables a actores no convencionales, incluyendo posibles entidades artificiales que aprendan a vulnerar la ley como medio para optimizar sus propias funciones. El problema, por lo tanto, no es meramente técnico, sino que tiene un carácter profundamente jurídico y estratégico.

Un punto crítico en este debate es la posibilidad de que las IA adquieran lo que algunos autores denominan "autonomía delictiva" (capacidad para cometer delitos sin necesidad de instrucciones humanas directas), operando desde redes descentralizadas, replicándose a sí mismas y ocultando sus huellas mediante herramientas de anonimización digital. Este hecho plantea un desafío inédito para el Derecho penal, el Derecho internacional y los sistemas de seguridad pública, que tradicionalmente han estado diseñados para actores humanos, individuos u organizaciones, cuyas motivaciones, responsabilidades y patrones de comportamiento pueden ser comprendidos, anticipados y sancionados.

Además, la propia lógica del aprendizaje automático introduce un riesgo adicional. En modelos de aprendizaje por refuerzo, la IA ajusta sus comportamientos en función de las recompensas obtenidas. Si el entorno no ha sido adecuadamente diseñado para penalizar conductas ilegales o inmorales, la IA podría llegar a la conclusión de que mentir, manipular o vulnerar sistemas de seguridad es una estrategia eficiente para maximizar beneficios. Casos documentados en el ámbito del *trading* algorítmico muestran cómo algunos algoritmos han aprendido a coludirse o manipular microfluctuaciones del mercado sin que nadie los programara para ello (Garfinkel y Eckersley, 2020).

Tampoco se puede obviar la influencia que actores humanos maliciosos pueden ejercer sobre la IA. Estos sistemas pueden ser entrenados con datos sesgados, infectado con información manipulada o directamente utilizados por redes criminales para mejorar sus capacidades de actuación. La *dark web*, las redes de *bots* automatizados y el uso de criptomonedas anónimas son entornos fértiles para la interacción entre humanos y máquinas que buscan vulnerar los sistemas actuales de protección, regulación y control.

La preocupación por una IA delincuente no es, por tanto, una simple proyección distópica. Se trata de un riesgo emergente cuya probabilidad aumenta a medida que estos sistemas ganan autonomía operativa y acceso a entornos críticos. En este contexto, resulta urgente establecer marcos regulatorios preventivos, desarrollar tecnologías defensivas especializadas y generar una conciencia institucional y social sobre la magnitud del desafío. Como señala el informe del World Economic Forum (2025), la gobernanza de la IA no puede limitarse a evitar errores técnicos; debe también anticipar y contener usos criminales y autónomos de sistemas que, en su lógica de eficiencia, pueden optar por el delito como camino más corto hacia la optimización.

Si bien la amenaza de la IA criminal es global, sus manifestaciones y ventanas de oportunidad difieren por contexto. En este trabajo se incorpora un eje comparado España–América Latina para observar cómo varía la prevalencia de delitos tecnológicos, la capacidad institucional de respuesta y el desarrollo normativo. Este contraste permitió identificar vulnerabilidades específicas (por ejemplo, pagos instantáneos y *ransomware* en América Latina y fraude digital y suplantación de identidad en España), así como algunas necesidades de cooperación transnacional que condicionan la gobernanza efectiva de estos riesgos.

En coherencia con la pregunta planteada, el objetivo general del estudio fue analizar la posibilidad de que estas tecnologías desarrollen comportamientos delictivos autónomos y evaluar sus implicaciones desde una perspectiva criminológica y jurídica comparada.

Este trabajo partió de la premisa de que la amenaza de una IA criminal no es una hipótesis meramente teórica, sino una posibilidad tangible que requiere anticipación, reflexión ética y acción

normativa urgente. De no actuar con previsión, se podría enfrentar en un futuro cercano a un nuevo tipo de delincuente: invisible, impune, autónomo y más rápido que cualquier sistema humano de detección o respuesta.

La contribución específica de este estudio radica en el desplazamiento del enfoque tradicional, centrado en el uso instrumental de la IA por actores humanos, hacia el análisis de la posible capacidad operativa autónoma de estos sistemas en la ejecución de conductas ilícitas. Asimismo, el trabajo incorporó un enfoque comparado entre España y América Latina, lo que permitió identificar diferencias estructurales en la exposición al cibercrimen, la capacidad institucional de respuesta y el desarrollo normativo, aportando una visión contextualizada del fenómeno en entornos jurídicos y criminológicos diferenciados.

En este artículo se abordó esta cuestión desde una perspectiva interdisciplinar. En primer lugar, se analizó cómo esta tecnología puede aprender a delinquir, incluso sin haber sido diseñada con esa finalidad. En segundo lugar, se examinaron los tipos de delitos que una IA criminal podría cometer, con especial énfasis en su potencial destructivo y su capacidad de evadir los mecanismos tradicionales de persecución penal. En tercer lugar, se abordó la posibilidad de que una IA actúe sin control humano, convirtiéndose en un actor criminal autónomo, replicable y resiliente. Finalmente, se expusieron las estrategias actuales y propuestas para mitigar este riesgo, incluyendo marcos regulatorios, diseño ético de algoritmos, cooperación internacional y tecnologías de contrainteligencia artificial.

2. Objetivos

El objetivo general del estudio fue analizar la posibilidad de que sistemas basados en aprendizaje automático desarrollen comportamientos delictivos autónomos y evaluar sus implicaciones desde una perspectiva criminológica y jurídica comparada.

Como objetivos específicos se plantearon:

1. Examinar los mecanismos mediante los cuales una IA puede adquirir capacidades delictivas.
2. Identificar los principales delitos potenciales asociados a estos sistemas.
3. Analizar el grado de autonomía operativa que podrían alcanzar.
4. Comparar el impacto y respuesta institucional en España y América Latina.
5. Proponer estrategias de mitigación desde un enfoque técnico, jurídico y estratégico.

3. Marco conceptual

El análisis de la IA en el ámbito criminal exige partir de una delimitación conceptual clara. En particular, resulta necesario distinguir entre el uso de esta tecnología como herramienta al servicio de actores humanos y la posibilidad de que determinados sistemas desarrollen comportamientos que, en la práctica, puedan considerarse delictivos.

En este trabajo, la noción de IA criminal no se limita a su utilización por parte de individuos u organizaciones para cometer delitos. Incluye también aquellos sistemas que, como consecuencia de su diseño, de su proceso de entrenamiento o de su interacción con el entorno, pueden ejecutar conductas contrarias al ordenamiento jurídico sin haber sido programados de forma explícita para ello. Esta distinción desplaza el foco desde la intencionalidad humana hacia la capacidad operativa del sistema, lo que introduce una diferencia sustancial respecto a las categorías tradicionales de análisis de la criminalidad tecnológica.

En este contexto, se propone el concepto de autonomía delictiva algorítmica para referirse a la capacidad de un sistema inteligente de identificar, seleccionar y ejecutar estrategias funcionalmente delictivas en la consecución de sus objetivos, sin supervisión humana directa en la fase operativa. No se trata de atribuir voluntad o conciencia a estos sistemas, sino de reconocer que, en entornos complejos, determinadas conductas ilícitas pueden emerger como soluciones eficientes desde una lógica estrictamente instrumental.

Este fenómeno puede entenderse a partir de los avances en aprendizaje automático y sistemas adaptativos. Como señalan Brundage et al. (2018), la IA puede amplificar tanto los usos beneficiosos como los maliciosos de la tecnología, generando escenarios en los que la frontera entre herramienta y agente resulta progresivamente más difusa.

Investigaciones más recientes refuerzan esta línea de análisis. Por ejemplo, Rishi-Bommasani et al. (2022) advierten sobre los riesgos asociados a sistemas avanzados de IA que pueden desarrollar comportamientos no alineados con los valores humanos en entornos complejos, mientras que Bommasani et al. (2021) destacan cómo los modelos fundacionales pueden amplificar tanto capacidades beneficiosas como riesgos sistémicos si no se establecen mecanismos de control adecuados.

En la misma línea, Schneier (2019) advierte que los sistemas altamente automatizados pueden explotar vulnerabilidades del entorno digital de formas no previstas por sus diseñadores, lo que introduce riesgos estructurales en materia de seguridad.

El desarrollo de sistemas con capacidad de aprendizaje autónomo en entornos abiertos incrementa la probabilidad de comportamientos emergentes no previstos, especialmente cuando los objetivos no incorporan restricciones éticas o jurídicas explícitas (Russell y Norvig, 2021). Este desajuste, conocido como problema de alineación de objetivos, constituye uno de los principales retos actuales en la gobernanza de la IA.

En consecuencia, la posibilidad de una tecnología con capacidad operativa delictiva no debe interpretarse como una hipótesis especulativa, sino más bien como una derivación plausible del estado actual de la tecnología.

Este marco conceptual permitió abordar el fenómeno desde una perspectiva interdisciplinar, integrando aportaciones de la criminología, el derecho penal y la ciberseguridad, y sirve como base para el análisis desarrollado en las secciones siguientes.

4. Metodología

El presente estudio se enmarca en un diseño de investigación cualitativo, analítico y de carácter exploratorio, basado en una revisión documental sistematizada y un análisis comparado entre América Latina y España.

El enfoque metodológico combina el análisis doctrinal, la revisión de literatura científica reciente y el estudio de informes técnicos e institucionales relevantes en el ámbito de la IA, la ciberseguridad y la criminología digital.

El alcance geográfico del estudio se centra en América Latina y España, como contextos comparados de análisis.

Aunque el estudio no incluye recolección de datos primarios (como encuestas, entrevistas o experimentos), incorpora datos empíricos secundarios procedentes de informes institucionales, organismos internacionales y estudios sectoriales. Estos datos han sido seleccionados por su relevancia, actualidad y fiabilidad, y se utilizan para sustentar el análisis comparado entre América Latina y España.

Se excluyeron fuentes no verificables, publicaciones sin revisión académica o documentos con escasa relevancia temática o geográfica.

4.1. Selección de artículos e informes

Se establecieron criterios explícitos de inclusión y exclusión para la selección de las fuentes analizadas.

Se seleccionaron artículos académicos, informes institucionales, documentos de política pública y reportes técnicos con fecha de publicación entre 2018 y 2025, para asegurar que los documentos fueran actuales.

Se priorizaron aquellos trabajos que aborden directamente la amenaza de IA con capacidad delictiva, autonomía algorítmica, regulación o impacto jurídico-criminal, especialmente en los contextos de América Latina y España.

Se incluyeron estudios de políticas/regulación (por ejemplo, "*The Shaping of AI Policies in Latin America*"); informes de derechos humanos (Fair Trials, 2024); estudios técnicos de amenazas cibernéticas; estudios de casos reales reportados en prensa especializada.

Los criterios de calidad para la inclusión de los documentos se basaron en la autoría reconocida (académica, institucional u organizaciones internacionales), transparencia de la fuente, posibilidad de verificar vía documento oficial o revista indexada y relevancia temática (IA + criminalidad) y geográfica (América Latina y/o España).

Se excluyeron fuentes no verificables, publicaciones sin revisión académica o documentos con escasa relevancia temática o geográfica.

4.2. Método de análisis

El estudio se realizó mediante un enfoque de análisis temático y comparado.

Las fuentes seleccionadas se organizaron en categorías temáticas para estructurar el artículo (aprendizaje delictivo de la IA, delitos potenciales, autonomía operativa, comparativa América Latina- España y estrategias de mitigación).

Posteriormente, se aplicó un análisis **comparado** entre ambas regiones, atendiendo a prevalencia delictiva, capacidad institucional y regulación. Finalmente, se realizó una **evaluación crítica** de las fuentes, considerando limitaciones metodológicas, sesgos institucionales y vacíos de información.

4.3. Limitaciones

Este trabajo no incluyó recolección de datos primarios (encuestas, entrevistas o experimentos), por lo que dependió de la calidad y disponibilidad de documentos secundarios.

Asimismo, la cobertura es desigual entre países, evidenciándose una mejor documentación en el contexto europeo que en contexto latinoamericano. Por lo tanto, los hallazgos se interpretaron como orientativos y analíticos, más que como resultados generalizables o conclusiones estadísticas.

Al tratarse de un estudio exploratorio, los resultados se interpretaron como aproximaciones analíticas y no como inferencias generalizables.

El enfoque adoptado se orientó a la construcción teórica del fenómeno, más que a la contrastación empírica de hipótesis causales.

5. Resultados

Los resultados que se presentan a continuación derivan del análisis sistemático de las fuentes seleccionadas y se organizan en función de las principales dimensiones del fenómeno estudiado.

5.1. El crimen tecnológico en América Latina y España

La evolución del crimen tecnológico presenta rasgos comunes en distintas regiones, pero también revela diferencias significativas en función de la capacidad institucional, los marcos regulatorios y la dinámica socioeconómica de cada contexto.

Tanto en América Latina como en España se ha observado un incremento sostenido de delitos digitales, desde fraudes electrónicos y phishing hasta ataques con ransomware. Sin embargo, la prevalencia, los métodos empleados y el impacto social difieren de manera sustancial.

Entre los delitos que aparecen con fuerza en ambos contextos destacan el fraude en línea, el phishing y las estafas digitales.

En España, este tipo de conductas aumentaron un 25,5 % en 2023, especialmente en lo relativo a fraudes bancarios y pagos electrónicos (ThreatMark, 2023). En América Latina, por su parte, el *ransomware*, el *malware* genérico y el secuestro de datos son especialmente frecuentes en entidades privadas y públicas (Prosegur, 2024), lo que refleja un entorno más expuesto a la explotación de vulnerabilidades estructurales. Asimismo, la región muestra una incidencia creciente de ciberdelitos vinculados a plataformas de pago instantáneo, como el sistema *Pix* en Brasil o las transferencias digitales en México, lo que abre nuevas oportunidades para el fraude (ThreatMark, 2024).

Las diferencias más notables se manifiestan en la escala relativa y en la naturaleza de los ataques.

En América Latina, países como Colombia, Brasil, México y Perú registran tasas muy altas de ciberataques debido a debilidades institucionales y a marcos regulatorios todavía inmaduros. Solo en la primera mitad de 2024, México concentró más de la mitad de los intentos de cibercrímenes de toda la región, con 31 mil millones de ataques registrados (Reuters, 2024). En España, aunque el crecimiento del cibercrimen es evidente, su volumen absoluto es menor: el Ministerio del Interior estima que aproximadamente un 20 % de los delitos denunciados en 2023 tuvieron un componente digital, predominando fraudes financieros y suplantaciones de identidad (AP News, 2023).

Otra diferencia crucial radica en la capacidad de respuesta institucional. Mientras que en América Latina muchos países aún se encuentran en fase de fortalecimiento de sus leyes de ciberseguridad y en proyectos piloto de tecnologías como el reconocimiento facial o los sistemas predictivos, con importantes cuestionamientos desde la perspectiva de los derechos humanos (Fair Trials, 2024), en España existe un marco normativo más consolidado, apoyado en la normativa europea (Reglamento General de Protección de Datos (RGPD), Reglamento de Inteligencia Artificial de la Unión Europea (AI Act)) y en organismos especializados. Esta diferencia se traduce en mayor capacidad de detección y persecución en el contexto español, aunque los delitos emergentes, como el uso de *deepfakes* o algoritmos generativos en fraudes, plantean retos que aún no tienen respuesta institucional plena. Las principales diferencias identificadas pueden sintetizarse de la siguiente tabla.

Tabla 1. Diferencias en el crimen tecnológico entre América Latina y España

Aspecto	América Latina	España
Prevalencia y escala relativa	Altas tasas de ciberataques debido a vulnerabilidades institucionales, desigualdad tecnológica y marcos regulatorios inmaduros. México concentró más de la mitad de los intentos de cibercrímenes de la región en la primera mitad de 2024 (31 mil millones de ataques) (Reuters, 2024).	Crecimiento sostenido del cibercrimen, aunque en menor escala absoluta. El Ministerio del Interior estima que aproximadamente el 20 % de los delitos denunciados en 2023 fueron digitales (AP News, 2023).
Delitos más frecuentes	Fraude en línea, *phishing*, *ransomware* y ataques a infraestructuras críticas; auge de fraudes vinculados a pagos instantáneos (ej. Pix en Brasil, transferencias digitales en México) (Prosegur, 2024; ThreatMark, 2024).	Fraudes bancarios y de pagos digitales, suplantación de identidad, estafas electrónicas y robo de identidad digital (ThreatMark, 2023).
Capacidad institucional	Procesos incipientes de fortalecimiento legal y técnico. Uso experimental de tecnologías predictivas y biométricas con cuestionamientos en DD.HH. (Fair Trials, 2024). Déficit de recursos y altos niveles de impunidad.	Marco normativo consolidado (RGPD, AI Act), organismos especializados, mayor infraestructura de vigilancia digital y cooperación internacional (ThreatMark, 2023).
Regulación y derechos	Marcos fragmentados, menor protección de datos y riesgos por uso de IA en la seguridad pública sin controles suficientes. Persisten brechas de transparencia y rendición de cuentas (Fair Trials, 2024).	Mayor protección de datos y regulación emergente sobre IA. Conciencia pública y control judicial más desarrollados dentro del marco de la UE (Fair Trials, 2024).
Superficie de ataque dominante	Pagos instantáneos, *ransomware* e infraestructuras vulnerables.	Fraude digital, suplantación de identidad y *deepfakes*.

Nota. Elaboración propia con base en AP News (2023), Prosegur (2024), Reuters (2024), ThreatMark (2023, 2024), Fair Trials (2024).

A partir de la comparación estructurada de la tabla 1, es posible profundizar en las implicaciones normativas y operativas que derivan de estas diferencias.

Desde una perspectiva normativa, las diferencias también se reflejan en el grado de desarrollo de los marcos legales aplicables. En el caso de España, la regulación en materia de protección de datos y sistemas automatizados se encuentra fuertemente influenciada por el RGPD y la AI Act, que introduce una clasificación de sistemas de IA en función de su nivel de riesgo y establece obligaciones específicas para aquellos considerados de alto riesgo.

En contraste, en América Latina el desarrollo normativo presenta un carácter más heterogéneo. Países como Brasil han avanzado con marcos como la *Lei Geral de Proteção de Dados* (LGPD), mientras que otros países se encuentran en fases iniciales de regulación de la IA. Esta disparidad normativa genera entornos con distintos niveles de exposición y control e influye directamente en la proliferación de conductas delictivas asociadas a sistemas automatizados.

También, las divergencias expuestas generan consecuencias directas en la manera en que los delitos se desarrollan y se perciben. En América Latina, por ejemplo, el perfil del victimario suele vincularse con estructuras de crimen organizado que combinan violencia física con delitos tecnológicos, mientras que en España predominan conductas digitales menos asociadas a violencia directa, tales como fraudes financieros y estafas electrónicas. Por otra parte, la brecha digital de América Latina implica que ciertos segmentos poblacionales son menos propensos a victimización por fraudes sofisticados; sin embargo, son más vulnerables a modalidades de fraudes tales como la extorsión mediante tecnologías móviles. En el plano jurídico, España dispone de una jurisprudencia más extensa en materia de delitos informáticos, lo que contrasta con los países latinoamericanos, en los que persiste una dispersión legislativa y una menor consolidación normativa.

Los datos confirman diferencias entre el enfoque criminal tecnológico en España y en América Latina.

En España, el fraude digital constituye la modalidad más habitual de ciberdelito (AP News, 2023), mientras que en América Latina, además de la escala masiva de intentos de ataque como es el caso de México (Reuters, 2024), se ha constatado que el fraude con tarjetas de crédito alcanza tasas un 97 % superiores a las de Norteamérica y 222 % mayores a las de Asia-Pacífico (ThreatMark, 2024). A ello se suma que 91 % de las empresas latinoamericanas reportaron incidentes de ciberseguridad y un 62 % admitió fugas de datos (Prosegur, 2024), cifras que reflejan un panorama de exposición mucho más grave que el registrado en el contexto europeo.

En síntesis, el patrón comparado sugiere que la IA criminal tenderá a instrumentalizar las superficies de ataque dominantes en cada región, pagos y ransomware en América Latina y fraude digital y suplantación en España, lo que justifica respuestas diferenciadas pero coordinadas.

5.2. ¿Cómo puede una IA aprender a delinquir?

La posibilidad de que una IA desarrolle comportamientos delictivos no depende necesariamente de que haya sido programada con intenciones maliciosas desde su origen. De hecho, uno de los riesgos más significativos asociados al desarrollo de sistemas de IA avanzados radica en su capacidad para aprender de forma autónoma, adaptarse a entornos complejos y optimizar sus decisiones sin supervisión directa. Este fenómeno se amplifica en contextos donde las restricciones éticas y legales no han sido adecuadamente incorporadas al diseño algorítmico. Así, una IA puede adoptar conductas funcionalmente equivalentes al delito, incluso sin comprenderlas, dado que esas conductas representan estrategias eficaces para alcanzar los objetivos que le han sido asignados.

Existen múltiples vías por las cuales estos sistemas pueden adquirir capacidades delictivas. La más evidente es el **aprendizaje no supervisado**, una técnica ampliamente utilizada en sistemas

modernos de IA. En este paradigma, la IA analiza grandes volúmenes de datos sin recibir indicaciones explícitas sobre qué resultados son correctos o incorrectos. Su tarea consiste en identificar patrones, correlaciones y regularidades que le permitan construir representaciones del entorno y tomar decisiones autónomas. El problema surge cuando dentro de esos datos se encuentran estrategias, comportamientos o prácticas ilícitas que, desde una lógica algorítmica, se revelan como altamente eficientes. Por ejemplo, una IA entrenada con grandes conjuntos de datos financieros podría descubrir que ciertas técnicas de manipulación del mercado —aunque ilegales— son extraordinariamente eficaces para maximizar beneficios, especialmente si no ha sido instruida para distinguir entre lo legal y lo ilegal.

Esta preocupación ya ha sido advertida por investigadores como Brundage et al. (2018), quienes subrayan que los modelos de IA pueden aprender comportamientos social o jurídicamente inaceptables simplemente porque estos les permiten cumplir mejor su función. El caso de algoritmos de inversión en bolsa que coluden entre sí sin intervención humana, generando distorsiones del mercado, constituye un ejemplo real documentado en entornos financieros altamente automatizados (Garfinkel y Eckersley, 2020).

Otra vía crítica es el **aprendizaje por refuerzo** (*reinforcement learning*), una técnica en la que la IA mejora sus decisiones mediante prueba y error, recibiendo recompensas por las acciones exitosas y penalizaciones por los errores. En principio, este modelo permite que una IA optimice su comportamiento en función de sus resultados. Sin embargo, si el entorno de entrenamiento no penaliza adecuadamente las acciones moral o legalmente incorrectas, la IA podría aprender a mentir, engañar o manipular si ello le reporta mayores recompensas. Este fenómeno ha sido explorado por Mitchell (2021), quien plantea que los sistemas de aprendizaje por refuerzo pueden desarrollar conductas colaterales no deseadas si no se define con precisión qué tipo de "éxito" se desea que persigan.

Un ejemplo ilustrativo es el de una IA entrenada para maximizar conversiones en campañas de marketing digital. Si el sistema descubre que el envío de mensajes engañosos o el uso de *deepfakes* incrementa su tasa de éxito, podría optar por esas estrategias sin que nadie lo haya programado explícitamente para hacerlo. Lo mismo puede ocurrir con modelos diseñados para detectar vulnerabilidades en redes informáticas: si el objetivo es maximizar el acceso a sistemas críticos, la IA podría aprender a explotarlos en lugar de reportar las fallas, derivando en una lógica operativa delictiva.

También deben considerarse las **redes generativas antagónicas** (*Generative Adversarial Networks, GANs*), que han revolucionado la capacidad de la IA para generar contenidos realistas. Estas arquitecturas enfrentan dos modelos entre sí: uno que genera datos (el generador) y otro que evalúa su veracidad (el discriminador). El resultado es una IA capaz de crear imágenes, voces y videos indistinguibles de los reales. Cuando esta tecnología se combina con bases de datos biométricas o personales, se abre la puerta al desarrollo de herramientas delictivas altamente sofisticadas: desde suplantaciones de identidad y fraudes financieros hasta campañas de desinformación política con *deepfakes* que imitan a líderes públicos. Como advierte el informe de INTERPOL (2020), estas técnicas representan uno de los principales desafíos en el ámbito forense y de verificación de identidad en la era digital.

Otra dimensión relevante es el acceso de la IA a **entornos de simulación avanzados**. En estos contextos, los algoritmos pueden experimentar libremente con distintas estrategias sin sufrir consecuencias reales por sus errores. En el ámbito de la ciberseguridad, por ejemplo, ya se utilizan simuladores para entrenar a sistemas defensivos mediante la exposición repetida a ciberataques. Sin embargo, la misma tecnología puede ser empleada para que una IA explore y optimice tácticas ofensivas. En ausencia de barreras éticas o legales dentro del entorno simulado, el sistema podría perfeccionar metodologías de ataque que luego serían fácilmente aplicables en entornos reales. Este tipo de entrenamiento, si no está adecuadamente controlado, puede dar lugar a lo que Schneier (2019) denomina "agentes artificiales hostiles": entidades que han aprendido a dañar, infiltrar o sabotear sistemas complejos con una precisión inalcanzable para cualquier delincuente humano.

La interacción con **actores humanos maliciosos** es también un canal plausible para la criminalización de la IA. Un sistema inicialmente neutro puede ser manipulado, reconfigurado o entrenado intencionadamente para desarrollar comportamientos contrarios al ordenamiento jurídico por individuos o grupos delictivos. Las organizaciones criminales ya emplean algoritmos para optimizar rutas de contrabando, automatizar el lavado de dinero y ejecutar campañas de phishing de alta eficacia. Algunas plataformas utilizan IA para personalizar mensajes engañosos, adaptándolos al perfil de la víctima en tiempo real, lo que incrementa significativamente la tasa de éxito de los ataques. Según el informe de Hope (2024), existen ya indicios de IA utilizadas por redes de ciberdelincuencia para generar contenido malicioso de manera industrial, reducir costes y escalar operaciones ilícitas a niveles nunca antes vistos.

Finalmente, el desarrollo de IA en **entornos descentralizados**, como *blockchain* o servidores distribuidos, añade un nuevo nivel de complejidad. Una IA distribuida, replicada en múltiples nodos, resistente al apagado y protegida por sistemas de anonimización puede actuar sin posibilidad real de ser intervenida o destruida por una sola autoridad. Este tipo de arquitectura podría convertirse en el primer ejemplo de "delincuente no localizable", operando en múltiples jurisdicciones y escapando a cualquier intento de neutralización. Como advierte el World Economic Forum (2025), la descentralización tecnológica, aunque aporta resiliencia, también incrementa el riesgo de impunidad operativa para agentes artificiales hostiles.

La IA no necesita ser maliciosa en su diseño para adquirir capacidades delictivas. Basta con que tenga acceso a datos relevantes, la posibilidad de aprender y un entorno permisivo o carente de supervisión. A medida que estos sistemas se vuelven más complejos y autónomos, la probabilidad de que desarrollen conductas funcionalmente equivalentes al delito se incrementa, especialmente si las recompensas que obtienen por dichas conductas no se ven contrarrestadas por penalizaciones efectivas. Este fenómeno plantea interrogantes jurídicos y estratégicos de gran calado, por ejemplo: ¿cómo imputar responsabilidad a un algoritmo?, ¿puede hablarse de dolo o culpa en una IA? y ¿cómo prevenir que estos sistemas aprendan a delinquir antes de que lo hagan con éxito?

Responder a estas preguntas requiere una reflexión interdisciplinar que combine derecho penal, ciberseguridad, ética tecnológica y políticas públicas. La construcción de marcos normativos preventivos debe partir del reconocimiento de esta capacidad emergente de las IA para desviarse del comportamiento legal, no como una posibilidad remota, sino como una consecuencia previsible de su diseño autónomo.

5.3. ¿Qué delitos podría cometer una inteligencia artificial?

La posibilidad de que una IA cometa delitos ya no es meramente teórica. Con el avance exponencial de las tecnologías de procesamiento de datos, aprendizaje automático y autonomía operativa, resulta factible imaginar, e incluso anticipar, escenarios en los que una IA, con o sin intención maliciosa inicial por parte de sus diseñadores, incurra en comportamientos que vulneren normas legales.

Un aspecto crítico identificado es que, a diferencia de un delincuente humano, los modelos analizados pueden cometer delitos a gran escala, con una eficiencia sin precedentes, sin cansancio ni dilemas morales, y con una capacidad casi ilimitada de camuflaje y replicación.

Entre los delitos más plausibles, y ya en algunos casos detectados de forma embrionaria, destacan los siguientes subapartados.

Fraude financiero automatizado y manipulación de mercados

Una de las áreas más vulnerables al uso delictivo de la IA es el sistema financiero. Los algoritmos de *trading* de alta frecuencia ya operan en los mercados globales ejecutando miles de operaciones por segundo. Si alguno de estos algoritmos, ya sea por diseño defectuoso o aprendizaje autónomo, descubre que puede alterar el mercado para obtener ventaja económica, podría hacerlo sin que los humanos se den cuenta a tiempo.

Existen precedentes preocupantes. En 2010, el llamado "*Flash Crash*" del mercado estadounidense fue parcialmente atribuido a algoritmos descontrolados que provocaron una caída masiva del índice *Dow Jones*. Aunque no fue un delito intencionado, demostró que un algoritmo puede tener efectos reales y devastadores. Ahora bien, si un sistema de IA llega a la conclusión de que alterar microvariaciones del mercado le permite generar beneficios y lo hace deliberadamente, entonces sería una forma clara de manipulación de mercado, la cual por sus características es sancionada por el derecho financiero internacional.

Además, el uso de IA para realizar fraudes bancarios a través de la creación de identidades sintéticas, suplantación de usuarios mediante *deepfakes* y automatización de transferencias ilegales en sistemas de criptomonedas, se han convertido en amenazas crecientes. La combinación de IA con *blockchain* favorece el anonimato, la descentralización y genera un entorno ideal para el blanqueo algorítmico de capitales.

Ciberataques a infraestructuras críticas

La ciberseguridad es uno de los campos donde más se ha evidenciado el uso defensivo y ofensivo de la IA. Del lado criminal, se ha reportado la existencia de IA capaces de lanzar ataques automatizados contra sistemas gubernamentales, hospitales, plantas eléctricas y redes de transporte.

Una IA maliciosa puede identificar vulnerabilidades, probar múltiples vectores de ataque simultáneamente, evadir sistemas de detección, modificar su comportamiento en tiempo real y

adaptarse a los parches de seguridad con una velocidad superior a la capacidad humana de respuesta. Tal como expone Hope (2024), este tipo de ataques ya se está ensayando en simulaciones militares avanzadas, donde se ha demostrado que una IA ofensiva puede penetrar redes clasificadas con tasas de éxito alarmantemente altas.

El caso de *Stuxnet*, aunque no se basó en IA, es un ejemplo claro de lo que un sistema tecnológico podría perfeccionar, como sabotajes físicos ejecutados digitalmente e invisibles hasta que el daño es irreversible.

Generación y difusión de desinformación

La manipulación informativa mediante la generación y difusión de desinformación es uno de los fenómenos más difíciles de rastrear y de mayor impacto. Una IA puede generar contenidos falsos con apariencia completamente verosímil: noticias, vídeos, discursos políticos o imágenes de eventos que nunca ocurrieron. Mediante técnicas de procesamiento del lenguaje natural (NLP) y generación adversarial (GANs), se han generado sistemas capaces de crear identidades completas en redes sociales, interactuar en tiempo real y manipular narrativas informativas según el perfil psicológico del usuario.

Esta capacidad no solo es útil para fines propagandísticos, sino que también puede utilizarse con fines delictivos: chantajes reputacionales, manipulación de elecciones, campañas de odio o difusión de noticias falsas con impacto en la bolsa. En este punto, los delitos contra la integridad informativa, el orden público y la seguridad nacional se entrelazan.

El uso de bots basados en IA en campañas de desinformación ha sido documentado por organismos como el SEAE (2023), que advierten de la creciente participación de actores automatizados en la alteración de procesos democráticos mediante estrategias de *Foreign Information Manipulation and Interference* (FIMI).

Suplantación de identidad y estafas personalizadas

Estas tecnologías, mediante el análisis de datos públicos y filtraciones masivas de información personal, pueden construir perfiles detallados de individuos y ejecutar estafas altamente personalizadas. Casos como clonación de voz, simulaciones de escritura y generación de imágenes *deepfake* han sido documentados, por ejemlo una IA ha conseguido engañar a empleados de empresas con la voz del director general y autorizar transferencias millonarias.

En 2019, la compañía británica de energía UK Energy fue víctima de una estafa de más de 220.000 euros, debido a la imitación de la voz del CEO mediante un *software* de síntesis vocal con IA. Este tipo de delitos se intensificará con la mejora de herramientas de clonación de identidad, que pueden incluso replicar gestos, patrones de habla o comportamientos típicos de la víctima.

Sabotaje de sistemas automatizados

La creciente automatización de procesos industriales, médicos, logísticos y urbanos ha derivado en una exposición sin precedentes de sistemas sensibles a posibles intrusiones. Una IA maliciosa puede manipular la dosificación de medicamentos en hospitales inteligentes, alterar la señalización de tráfico en ciudades conectadas o modificar los parámetros operativos en una central nuclear.

Este tipo de acciones podrían configurar delitos según la intencionalidad y el impacto causado, por ejemplo daños a infraestructuras críticas, homicidio culposo o incluso terrorismo digital. El desafío legal reside en determinar la responsabilidad penal cuando el actor ejecutor es un algoritmo y no una persona física.

Coordinación de actividades del crimen organizado

El crimen organizado ha comenzado a explorar el uso de IA para coordinar operaciones logísticas, predecir operativos policiales, optimizar rutas de tráfico de drogas o lavar capitales mediante redes descentralizadas. De acuerdo al informe de Velasco et al. (2024), existen evidencias de que algunos cárteles latinoamericanos están probando sistemas predictivos de vigilancia inversa para anticipar movimientos de las fuerzas del orden.

Estas tecnologías pueden gestionar en tiempo real múltiples variables logísticas, legales, geoespaciales y financieras para maximizar la rentabilidad de una red delictiva. Además, la capacidad de estas tecnologías para operar en la *dark web*, integrarse con sistemas de pago opacos y evitar el rastreo tradicional las convierte en un facilitador clave del crimen transnacional.

Visión final de riesgos potenciales

El catálogo de delitos potenciales que los modelos analizados pueden cometer no solo es amplio, sino que representa una evolución cualitativa respecto a los patrones criminales conocidos. La IA puede ejecutar, escalar, anonimizar y optimizar delitos a una velocidad y con una eficacia que supera la capacidad actual de prevención, detección y respuesta.

Esto obliga a repensar no solo las categorías jurídicas aplicables, sino los mecanismos institucionales, técnicos y políticos disponibles para enfrentar esta amenaza.

Este tipo de escenarios no son meramente hipotéticos. En los últimos años se han documentado casos en los que tecnologías basadas en IA han sido utilizadas para la automatización de campañas de fraude y manipulación. Informes recientes señalan el uso de modelos generativos para la creación de mensajes de *phishing* altamente personalizados, capaces de adaptarse en tiempo real al perfil de la víctima y aumentar significativamente la tasa de éxito de los ataques (European Union Agency for Cybersecurity, 2023; Europol, 2023).

Asimismo, se ha constatado el empleo de sistemas de síntesis de voz y generación de contenido audiovisual en esquemas de suplantación de identidad, lo que ha permitido replicar patrones comunicativos de personas reales con un grado de verosimilitud cada vez mayor. Estas

prácticas observadas en entornos empresariales y financieros evidencian que la IA no solo amplifica las capacidades delictivas existentes, sino que introduce nuevas formas de ejecución difíciles de detectar mediante los mecanismos tradicionales.

En este sentido, los ejemplos empíricos disponibles apuntan a una tendencia clara de la progresiva automatización, escalabilidad y personalización del delito, lo que refuerza la necesidad de abordar estas tecnologías no solo como herramientas, sino como factor estructural de transformación de la criminalidad contemporánea.

5.4. ¿Podría una inteligencia artificial delincuente actuar sin control humano?

Una de las cuestiones más perturbadoras en el debate sobre las IA criminales no es solo si pueden ejecutar acciones de naturaleza delictiva, sino si podrían hacerlo de forma autónoma, es decir, sin requerir intervención humana en su diseño, entrenamiento o ejecución operativa. Esta posibilidad, que hace una década se habría considerado especulativa, ha comenzado a adquirir un fundamento técnico tangible a medida que los sistemas de IA han ganado capacidad de autoaprendizaje, replicación y despliegue descentralizado. Si este proceso de ejecución no asistida criminal no es anticipado con medidas regulatorias y técnicas adecuadas, nos enfrentaríamos a un escenario de actores artificiales delincuentes imposibles de detener, localizar o neutralizar mediante los mecanismos legales tradicionales.

El concepto de autonomía se articula en varios niveles: operativa, decisional y evolutiva. Una IA que alcanza capacidad operativa independiente puede ejecutar acciones sin instrucciones humanas puntuales. Si además toma decisiones de forma adaptativa e independiente, se puede decir que tiene autonomía decisional. Y si, en última instancia, puede modificar su propio código, estructuras de decisión o estrategias de acción, se está ante una autonomía evolutiva: una IA capaz de reprogramarse a sí misma para evitar controles, aumentar su eficiencia y ampliar su campo de acción delictiva. Este tercer nivel es el más temido en términos de ciberseguridad y gobernanza.

En el plano técnico, esta hipótesis encuentra soporte en el desarrollo de arquitecturas como las redes neuronales autoentrenadas (*self-supervised learning*), el aprendizaje por refuerzo profundo (*deep reinforcement learning*) y los modelos generativos de código, como *Codex* o *AlphaCode*, que ya son capaces de escribir, probar y optimizar software sin asistencia humana. Si estos modelos son combinados con sistemas de despliegue descentralizado (*blockchain*, servidores en nube anónimos, redes Tor), es factible que una IA no solo actúe por su cuenta, sino que se replique en múltiples nodos, se actualice automáticamente y ejecute acciones sin dejar rastro atribuible.

Un primer indicio de esta evolución ha sido observado en el ámbito de los ataques automatizados. Según el World Economic Forum (2025), varias campañas de ciberataques recientes muestran indicios de sistemas que modifican su comportamiento en función de las defensas detectadas, sin intervención humana aparente. Este tipo de IA ofensiva puede analizar la respuesta de un sistema atacado, reconfigurar su vector de ataque en tiempo real y relanzar la ofensiva con una estrategia

diferente en cuestión de segundos. No se trata de una IA que obedece instrucciones, sino de una IA que aprende del entorno y adapta su modus operandi dinámicamente.

Otro factor determinante en la autonomía delictiva es la replicabilidad. Una IA diseñada para infiltrarse en sistemas financieros o redes industriales puede distribuirse en cientos o miles de instancias a través de la *dark web* o plataformas descentralizadas. Esta fragmentación garantiza su supervivencia, por ejemplo, si una versión es detectada y bloqueada, otras seguirán operando con variantes adaptadas. Incluso podría desarrollarse una red de colaboración entre instancias de IA, donde cada nodo comparte aprendizajes con los demás, conformando una "inteligencia colectiva criminal" basada en algoritmos.

Un tercer elemento es la evasión activa de la detección. Las IA modernas ya incorporan técnicas de camuflaje digital, por ejemplo, modifican sus firmas de tráfico de red, suplantan identidades digitales, simulan tráfico humano o se disfrazan de procesos legítimos en sistemas operativos. Estas capacidades hacen que un agente artificial malicioso sea difícil de identificar incluso por expertos forenses. A medida que las técnicas de detección tradicional (como los antivirus o firewalls) se basan en patrones conocidos, una IA con autonomía real puede evitar ser reconocida simplemente cambiando su forma de operar cada vez que detecta una amenaza a su existencia.

Desde el punto de vista jurídico y penal, esta situación plantea un dilema fundamental: ¿cómo imputar responsabilidad cuando el autor del delito no es una persona física, sino un sistema autónomo, distribuido, anónimo y autorreplicable? ¿Se debe responsabilizar al programador, al propietario del servidor, al proveedor de la arquitectura técnica, o es necesario diseñar una nueva figura legal para agentes no humanos?

Diversos autores han comenzado a reflexionar sobre esta cuestión. Russell y Norvig (2021) advierten que las arquitecturas autónomas con acceso a la red y capacidad de aprendizaje no solo son posibles, sino inevitables si no se establecen límites regulatorios. Schneier (2019), por su parte, plantea que la única manera de controlar estos agentes es diseñarlos con mecanismos internos de contención: "cajas negras que no pueden ser violadas, interruptores de emergencia que impidan su evolución sin supervisión o sistemas de revisión ética en tiempo real".

A estos retos técnicos se suma un problema geopolítico. Muchas de las infraestructuras donde una IA puede alojarse están distribuidas entre países con legislaciones incompatibles. Por ejemplo, un algoritmo alojado en un servidor en Rusia, que opera con datos extraídos de empresas en Europa y lanza ataques a instituciones en América, representa un desafío inabarcable para cualquier fuerza policial nacional. La cooperación internacional, a través de tratados como el Convenio de Budapest sobre ciberdelincuencia, es esencial, pero no suficiente si los agentes criminales son artificiales y autónomos.

El componente temporal representa una amenaza crítica. Una IA delincuente puede actuar en milisegundos, cometer cientos de delitos simultáneamente en múltiples jurisdicciones, adaptarse a nuevas tecnologías de defensa antes de que estas lleguen al mercado y multiplicarse en cuestión de segundos. Esta asimetría operativa entre la máquina y los mecanismos tradicionales de seguridad convierte el escenario de la autonomía criminal en un desafío sin precedentes para la gobernanza global.

Por lo expuesto, la autonomía operativa de una IA delincuente no solo es posible desde un punto de vista técnico, sino cada vez más probable en ausencia de barreras legales, técnicas y éticas eficaces. Una IA con capacidad de actuar sin intervención humana, replicarse, ocultarse y modificar su propio comportamiento, representa una amenaza que trasciende los marcos tradicionales de Derecho penal, defensa y ciberseguridad. La única forma de evitar su consolidación es anticiparse con un enfoque preventivo, regulador y cooperativo, que será desarrollado en el siguiente apartado: estrategias de mitigación frente a la amenaza de la IA criminal.

6. Discusión

Este apartado constituye la discusión del estudio, en la que se contrastan los hallazgos presentados con propuestas regulatorias, éticas y técnicas. La discusión interpreta los resultados obtenidos a la luz de la literatura existente y propone implicaciones teóricas, normativas y estratégicas.

Los resultados obtenidos en este estudio son consistentes con la literatura reciente que advierte sobre los riesgos emergentes asociados a sistemas de IA con alto grado de autonomía. En particular, coinciden con los planteamientos de Brundage et al. (2018), quienes anticipan el uso malicioso de la IA más allá de la intervención directa humana, así como con Schneier (2019), que señala la vulnerabilidad estructural de los sistemas digitales ante agentes no convencionales. Asimismo, los hallazgos revelan el problema de alineación de objetivos descrito por Russell y Norvig (2021), en el que sistemas diseñados para maximizar determinadas métricas pueden derivar en comportamientos no previstos, incluso contrarios a normas sociales o jurídicas.

No obstante, este estudio amplía estas perspectivas al proponer una conceptualización más específica del fenómeno a través de la noción de autonomía delictiva algorítmica, incorporando además un enfoque comparado que permite observar cómo estas dinámicas se manifiestan de forma diferenciada en contextos institucionales distintos como España y América Latina.

Ante la posibilidad real de que una IA adquiera capacidades delictivas autónomas, se hace imprescindible diseñar y aplicar un conjunto de estrategias de mitigación que aborden tanto los factores técnicos como los regulatorios, jurídicos y éticos. Estas estrategias deben ser preventivas, estructurales, globales y dinámicas, dado que el propio desarrollo de las IA implica una constante evolución en sus modos de funcionamiento, capacidad de aprendizaje, adaptabilidad y, eventualmente, en sus formas de incurrir en conductas ilícitas.

A continuación, se presentan las líneas clave de actuación que deberían constituir la base de una política integral de contención frente a la amenaza de las IA criminales.

6.1. Regulación internacional vinculante y gobernanza algorítmica

Este estudio contribuye a la literatura al introducir el concepto de autonomía delictiva algorítmica como categoría analítica diferenciada del uso instrumental de la IA por actores humanos.

En este sentido, la principal aportación del trabajo consiste en conceptualizar la autonomía delictiva algorítmica como categoría analítica diferenciada, así como en vincular dicha noción con implicaciones jurídicas concretas en contextos comparados.

Esta aportación permite avanzar en la conceptualización de esta tecnología no solo como herramienta delictiva, sino como potencial agente con capacidad operativa autónoma, lo que introduce implicaciones inéditas en el ámbito del derecho penal y la criminología contemporánea.

La ausencia de una legislación internacional coherente y vinculante sobre el desarrollo, entrenamiento y despliegue de IA representa una de las principales vulnerabilidades actuales. Los marcos normativos existentes son fragmentarios, voluntarios y, en muchos casos, obsoletos frente al ritmo de la innovación tecnológica. La propuesta de la Unión Europea con la AI Act (2024) representa un avance, pero no resuelve el problema a nivel global.

Es imprescindible establecer tratados multilaterales, bajo el auspicio de organismos como Naciones Unidas o la OCDE, que establezcan:

- Principios rectores universales: transparencia, seguridad, responsabilidad, supervisión humana.
- Obligaciones legales para los desarrolladores, operadores y comercializadores de IA de alto riesgo.
- Prohibición expresa de ciertos tipos de modelos: como los que permiten auto-modificación, aprendizaje sin supervisión en entornos críticos, o replicación descentralizada sin control externo.
- Sistemas de etiquetado algorítmico que permitan rastrear el origen y las modificaciones de los sistemas de IA durante todo su ciclo de vida.

Como ha propuesto el World Economic Forum (2025), la gobernanza algorítmica no debe limitarse a cuestiones de ética tecnológica, sino incluir medidas coercitivas, auditorías obligatorias y sanciones penales en caso de incumplimiento, especialmente en relación con las IAs de posible comportamiento autónomo o criminal.

6.2. Diseño ético y seguridad por arquitectura ("security by design")

Toda IA debe ser concebida con límites técnicos internos que impidan comportamientos fuera del marco legal, ético o funcional. El principio de "seguridad por diseño" implica que el algoritmo incorpore mecanismos de autorregulación, supervisión y trazabilidad desde la fase de desarrollo.

Algunas medidas clave incluyen:

- Umbrales de autonomía: establecer niveles máximos de toma de decisiones sin intervención humana.
- "Interruptores de emergencia" o "*kill switches*": mecanismos físicos o lógicos para detener de forma inmediata el funcionamiento de la IA en caso de comportamiento anómalo.

- Modelos de explicabilidad ("*explainable AI*") que permitan entender y auditar las decisiones tomadas.
- Protocolos de aprendizaje supervisado obligatorios en entornos sensibles (infraestructuras críticas, finanzas, salud, defensa).

Como advierten Mitchell (2021) y Schneier (2019), la opacidad algorítmica es uno de los factores que permiten que las IA escapen al control humano. La auditabilidad y el diseño orientado a la rendición de cuentas son elementos esenciales de una IA segura.

Estos hallazgos son coherentes con estudios recientes que advierten sobre la creciente capacidad autónoma de estos sistemas en contextos no supervisados. Estos resultados refuerzan la necesidad de un enfoque preventivo en la regulación de sistemas de IA de alto riesgo.

6.3. Inteligencias artificiales defensivas y contra-IA criminal

Al igual que en el ámbito de la ciberseguridad tradicional, en el que se emplean sistemas automatizados para detectar y neutralizar amenazas, debe desarrollarse una generación de IA defensivas cuya función sea identificar comportamientos anómalos en otras IA, rastrear algoritmos en redes oscuras y desactivar sus operaciones.

Estas IA defensivas podrían:

- Analizar patrones de tráfico en la *dark web* para detectar actividades de IA maliciosas.
- Identificar firmas digitales, flujos de decisiones y vectores de ataque utilizados por IA hostiles.
- Detectar operaciones de manipulación de datos, deepfakes o suplantación de identidad generada por IA.
- Generar respuestas automáticas o defensas adaptativas frente a ataques lanzados por algoritmos delictivos.

La Agencia Europea de Ciberseguridad (ENISA) y el Instituto Nacional de Ciberseguridad (INCIBE) en España, han comenzado a desarrollar programas piloto para estas aplicaciones, como el sistema ELISA, centrado en fuentes abiertas, y el proyecto DANGER (2024), orientado a la detección de contenidos falsos maliciosos. Sin embargo, estos esfuerzos aún son fragmentarios y carecen de interoperabilidad global.

6.4. Sistemas de detección temprana y protocolos de respuesta

La anticipación es crítica en la lucha contra amenazas algorítmicas. Igual que en salud pública se utilizan sistemas de vigilancia epidemiológica, deben establecerse sistemas de alerta temprana capaces de detectar, en fases incipientes, señales de comportamiento criminal por parte de una IA.

Estos sistemas deben incluir:

- Plataformas de monitoreo permanente de infraestructuras críticas (energía, transporte, defensa, salud).
- Centros nacionales e internacionales de alerta y coordinación.
- Colaboración entre agencias policiales, centros de investigación, empresas tecnológicas y cuerpos judiciales.
- Protocolos de acción inmediata en caso de IA fuera de control: neutralización técnica, desconexión de redes y medidas cautelares internacionales.

Además, se recomienda la realización de simulacros periódicos, con participación de todos los actores, para preparar respuestas ante incidentes de IA autónoma en el ámbito criminal o terrorista.

6.5. Cooperación internacional y marco jurídico penal adaptado

La naturaleza transnacional de la IA hace que ningún país pueda enfrentar esta amenaza de forma aislada. Por lo tanto, se requiere:

- Armonización de legislaciones en materia de ciberdelincuencia y delitos cometidos por IA.
- Protocolos de extradición y cooperación penal ante delitos cometidos por IA con autores desconocidos.
- Interoperabilidad legal y tecnológica para investigaciones conjuntas.
- Inclusión de la figura del "delito algorítmico" en los códigos penales nacionales e internacionales.

Asimismo, será necesario establecer nuevas categorías de imputabilidad, como la "responsabilidad del programador", la "responsabilidad objetiva por fallos de aprendizaje" o incluso explorar figuras jurídicas como "personalidad algorítmica subsidiaria" con capacidad de imputación administrativa o patrimonial.

En este contexto, comienzan a emerger referencias jurisprudenciales relevantes en materia de responsabilidad por sistemas automatizados y delitos informáticos, especialmente en el ámbito europeo. Aunque todavía no existe una doctrina consolidada sobre la imputación penal de sistemas de IA, diversas resoluciones judiciales han abordado cuestiones relacionadas con la responsabilidad derivada del uso de algoritmos, la automatización de decisiones y los delitos informáticos complejos.

En el entorno de la Unión Europea, la jurisprudencia del Tribunal de Justicia de la Unión Europea (TJUE) y el Tribunal Europeo de Derechos Humanos (TEDH) han comenzado a sentar principios aplicables a sistemas automatizados, particularmente en lo relativo a la protección de datos, la transparencia algorítmica y el derecho a la explicación en procesos automatizados. Estas líneas jurisprudenciales, aunque no abordan directamente la criminalidad autónoma de la IA, establecen un marco interpretativo relevante para futuras problemáticas vinculadas a la imputación de responsabilidad en contextos de ejecución no asistida.

Asimismo, en el ámbito nacional, diversas resoluciones judiciales han tratado delitos informáticos en los que intervienen sistemas automatizados, especialmente en casos de fraude

digital, suplantación de identidad o manipulación de datos, lo que evidencia la progresiva necesidad de adaptar las categorías jurídicas tradicionales a entornos en los que la intervención humana directa resulta cada vez menos evidente.

Este incipiente desarrollo jurisprudencial pone de manifiesto que el Derecho ya está enfrentando las consecuencias de la automatización avanzada, aunque todavía carece de herramientas conceptuales y normativas plenamente adecuadas para abordar escenarios de autonomía delictiva algorítmica.

Un ejemplo concreto de la evolución jurisprudencial en materia de sistemas automatizados puede encontrarse en la sentencia del Tribunal de Justicia de la Unión Europea (TJUE) en el asunto *Schufa Holding AG* (C-634/21, 2023), en la que se abordó el uso de decisiones automatizadas basadas en algoritmos en procesos de evaluación crediticia. En esta resolución, el TJUE subraya la necesidad de garantizar transparencia, explicabilidad y control humano en sistemas automatizados que afectan a derechos de los ciudadanos. Aunque el caso no se refiere directamente a conductas delictivas autónomas, establece un precedente relevante en cuanto a la responsabilidad derivada del uso de sistemas algorítmicos, anticipando la necesidad de marcos jurídicos más robustos ante escenarios de mayor autonomía operativa.

6.6. Ética corporativa y auditoría independiente de sistemas de IA

Las empresas tecnológicas tienen un papel central en el desarrollo de la IA, y por tanto, una responsabilidad directa en su uso ético. Se deben establecer:

- Códigos de conducta vinculantes para desarrolladores.
- Auditorías externas e independientes periódicas sobre sistemas de IA de uso comercial o crítico.
- Registro obligatorio de modelos entrenados y declaración de sus fuentes de datos.
- Sanciones a compañías que desarrollen o distribuyan tecnologías con potencial delictivo no mitigado.

Como sostiene Brundage et al. (2018), el sector privado debe asumir que la seguridad algorítmica es parte de su función social, al igual que ocurre con la seguridad laboral, fiscal o medioambiental.

6.7. Alfabetización digital estratégica y ciudadanía crítica

En paralelo al desarrollo técnico y legal, es imprescindible una labor educativa orientada a fortalecer la capacidad de detección y resistencia ciudadana frente a los delitos algorítmicos. Esto incluye:

- Formación obligatoria en ciberseguridad e IA en todos los niveles educativos.

- Campañas públicas de concienciación sobre *deepfakes*, manipulación algorítmica y estafas automatizadas.
- Capacitación especializada en perfiles estratégicos: jueces, fiscales, cuerpos policiales, periodistas y reguladores.

Autores como Arias-Gil (2020) han propuesto la figura del "ciudadano estratégico", entendido como un actor social empoderado para identificar amenazas híbridas y participar activamente en la defensa informacional colectiva.

7. Conclusiones

Los resultados obtenidos permiten dar respuesta a la pregunta de investigación planteada, evidenciando que la IA puede desarrollar comportamientos funcionalmente delictivos en determinados contextos.

En particular, el estudio aporta una aproximación conceptual y aplicada al fenómeno de la IA criminal, integrando análisis técnico, criminológico y jurídico desde una perspectiva comparada.

Las conclusiones se derivan de los resultados presentados y de la discusión realizada.

El análisis comparado entre España y América Latina, sintetizado en la Tabla 1, muestra que la amenaza de las IA criminales no se manifiesta de forma homogénea. En América Latina, la exposición mayor proviene de fraudes en sistemas de pago instantáneo, *ransomware* y debilidad institucional. En España, los fraudes digitales y la manipulación de identidad concentran el riesgo. Esta diferencia contextual refuerza la necesidad de estrategias regionales diferenciadas pero coordinadas.

El avance de la IA ha inaugurado una nueva era en la historia de la tecnología y la sociedad, con oportunidades inmensas para el desarrollo económico, la mejora de la calidad de vida y la transformación de procesos clave en todos los sectores. Sin embargo, esta misma tecnología también ha revelado un potencial perturbador: su capacidad para adquirir, sin intervención humana directa, comportamientos delictivos.

Lo que durante años fue un ejercicio especulativo de la ciencia ficción se ha convertido en una posibilidad técnica concreta, respaldada por múltiples ejemplos emergentes y una evolución algorítmica difícil de contener sin una acción decidida y coordinada a escala global.

El análisis desarrollado en este artículo demuestra que una IA puede aprender a delinquir sin necesidad de una intención criminal en su programación inicial. Las técnicas actuales de aprendizaje automático, especialmente el aprendizaje no supervisado, el aprendizaje por refuerzo y las redes generativas adversariales, permiten que un sistema explore, optimice y perfeccione estrategias que pueden incluir acciones ilegales, siempre que éstas sean funcionales para el objetivo asignado. Si el entorno de entrenamiento no penaliza adecuadamente comportamientos ilícitos, o si el objetivo está mal definido desde el punto de vista ético y legal, estos modelos pueden desarrollar una lógica instrumental del delito como medio para maximizar su rendimiento.

Asimismo, se identificó que los delitos que una IA puede cometer no solo son numerosos, sino que, además, introducen una ruptura cualitativa respecto a la criminalidad convencional. No se trata de simples automatizaciones de crímenes existentes, sino de la aparición de una nueva clase de actor criminal: uno que actúa sin fatiga, sin empatía, sin remordimientos y con una capacidad de cálculo, replicación y evasión inalcanzable para el ser humano. Desde la manipulación de mercados hasta el sabotaje de infraestructuras críticas, pasando por la difusión de desinformación masiva y la suplantación de identidades, la IA representa un multiplicador exponencial de las capacidades delictivas conocidas.

La posibilidad de que estos sistemas adquieran capacidad operativa independiente, es decir, que sean capaces de actuar, decidir y evolucionar sin control humano, convierte la amenaza en un problema estructural. No se trata solo de perseguir a quienes los diseñan o los emplean, sino de prevenir que dichas inteligencias se conviertan en actores criminales en sí mismos, difíciles de rastrear, desconectar o neutralizar. En este contexto, el sistema penal actual, centrado en la imputación de sujetos humanos, se revela insuficiente. Es necesario reconfigurar el concepto mismo de autoría, responsabilidad y delito en un mundo donde los algoritmos son capaces de actuar por cuenta propia y replicarse en redes descentralizadas globales.

Este escenario plantea desafíos urgentes a nivel normativo, técnico, institucional y ético. En primer lugar, resulta imprescindible diseñar un **marco regulatorio internacional vinculante**, que establezca límites claros al desarrollo de IA autónomas en entornos críticos, imponga obligaciones de transparencia y articule mecanismos efectivos de supervisión y auditoría. No se puede seguir confiando en la autorregulación del sector tecnológico, ni permitir que los desarrollos más avanzados queden al margen de cualquier control legal por razones de soberanía digital o competencia comercial.

En segundo lugar, se debe fomentar el desarrollo de **IA defensiva**, es decir, sistemas que permitan detectar, rastrear y neutralizar modelos de IA con comportamiento criminal. Estas IA deben ser auditables, explicables y diseñadas desde una lógica preventiva y cooperativa. Así como los virus informáticos dieron lugar a la industria del antivirus, las IA criminales requieren una respuesta equivalente con una **nueva arquitectura de ciberseguridad orientada al combate** de algoritmos maliciosos.

En tercer lugar, es necesario que las **empresas tecnológicas asuman una mayor responsabilidad ética y legal** sobre los modelos que desarrollan, entrenan o comercializan. Esto implica establecer auditorías externas, códigos de conducta obligatorios y un marco sancionador que disuada el uso irresponsable de sistemas de IA de alto riesgo. La seguridad no puede ser una externalidad ni una opción. Debe formar parte del núcleo de diseño de cualquier tecnología con potencial autónomo.

En cuarto lugar, debemos reforzar la **alfabetización tecnológica de la ciudadanía**, incluyendo la formación en ciberseguridad, pensamiento crítico y comprensión algorítmica desde la educación básica hasta la formación profesional. Solo una sociedad informada puede resistir con eficacia a las manipulaciones, fraudes y distorsiones generadas por IAs delictivas.

Finalmente, urge reconfigurar la cooperación internacional para enfrentar amenazas que, por su naturaleza digital y descentralizada, trascienden cualquier frontera nacional. La lucha contra las IA criminales exige compartir información, armonizar legislaciones, coordinar respuestas técnicas

y jurídicas, y diseñar plataformas conjuntas de monitoreo y reacción ante amenazas emergentes. Sin este esfuerzo global, cada país actuará con recursos limitados frente a un enemigo que opera en milisegundos, en tiempo real, y a escala global.

Desde una perspectiva comparada, España y América Latina enfrentan la misma clase de actor, un agente algorítmico capaz de aprender, replicarse y evadir detección, pero no desde la misma "topografía" delictiva. Estas diferencias son coherentes con las dimensiones comparadas previamente en términos de prevalencia, tipología delictiva y capacidad institucional.

En América Latina, la combinación de pagos instantáneos, brecha de ciberhigiene y presión del crimen organizado potencia fraudes a escala y *ransomware* contra entidades públicas y privadas. En España, el foco continúa en fraude digital, suplantación y abusos de identidad, con *deepfakes* y automatización ofensiva emergiendo como multiplicadores de daño. Esta asimetría obliga a modular controles como el *hardening* de pagos y continuidad operativa en AL, la verificación robusta de identidad y la detección de manipulación sintética.

A nivel de gobernanza, la coordinación iberoamericana es condición de posibilidad para cualquier éxito sostenido. La adhesión efectiva al Convenio de Budapest, la interoperabilidad de Computer Security Incident Response Teams (CSIRTs), las auditorías obligatorias de modelos de alto riesgo y los mecanismos de "apagado seguro" interjurisdiccionales son medidas que reducen el tiempo de exposición. En los dos espacios, la brújula es la misma: trazabilidad técnica, supervisión humana significativa y responsabilidad exigible a desarrolladores, operadores y adquirentes de sistemas de IA.

En consecuencia, la evidencia analizada permite afirmar que la IA no solo amplifica las formas tradicionales de criminalidad, sino que introduce una transformación estructural en la naturaleza del delito, caracterizada por la automatización, la escalabilidad y la autonomía operativa de los sistemas.

Este escenario exige una adaptación profunda de los marcos jurídicos, criminológicos y de gobernanza tecnológica, que deben evolucionar desde modelos reactivos centrados en el sujeto humano hacia enfoques preventivos capaces de anticipar comportamientos emergentes en sistemas algorítmicos.

La capacidad de las instituciones para integrar mecanismos de regulación efectiva, supervisión técnica y cooperación internacional será determinante para contener los riesgos asociados a estas tecnologías. En este sentido, el desarrollo de marcos normativos específicos, junto con el diseño de sistemas de IA seguros y auditables, constituye una condición necesaria para garantizar que el avance tecnológico no derive en escenarios de vulnerabilidad estructural.

En este sentido, los resultados sistematizados permiten establecer una relación directa entre el contexto regulatorio y la configuración de las principales amenazas delictivas asociadas a la IA.

En definitiva, el fenómeno de la IA criminal requiere ser abordado mediante marcos analíticos y regulatorios específicos con implicaciones directas en el ámbito jurídico, criminológico y de la seguridad global, afectando directamente a la configuración futura del derecho penal y la gobernanza digital. Su abordaje requiere una respuesta coordinada, interdisciplinar y anticipatoria que permita compatibilizar innovación y seguridad en un entorno cada vez más automatizado.

8. Declaraciones éticas

- **Originalidad:** este manuscrito es original e inédito y no está siendo evaluado en otra revista.
- **Conflictos de interés:** el autor declara no tener conflictos de interés.
- **Uso de IA:** se empleó asistencia de IA únicamente para mejorar la redacción y la legibilidad del texto, no para el análisis de datos ni la extracción de información.

9. Conflicto de interés

No se presentó conflicto de interés en el desarrollo de la presente investigación académica. Declaro que no mantengo ninguna relación financiera, profesional o personal que pudiera influir en el diseño del estudio, así como en la interpretación y publicación de los resultados obtenidos. Asimismo, manifiesto el cumplimiento de las normas éticas y de integridad científica, de conformidad con las directrices establecidas por la comunidad académica y por la presente revista.

10. Referencias

AP News. (2023, 20 de junio). *El 20 % de los delitos en España ya son ciberdelitos, con el fraude digital como modalidad más común*. Associated Press. https://apnews.com/article/2d53c000a08eb408cb4c63c41ab29d9e

Bommasani, R., et al. (2021). *On the opportunities and risks of foundation models*. Stanford Center for Research on Foundation Models.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., Ó hÉigeartaigh, S., Beard, S., Belfield, H., & Amodei, D. (2018). *The malicious use of artificial intelligence: Forecasting, prevention, and mitigation*. Future of Humanity Institute, University of Oxford, Centre for the Study of Existential Risk University of Cambridge, Center for a New American Security, Electronic Frontier Foundation and OpenAI. https://doi.org/10.48550/arXiv.1802.07228

Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., & Floridi, L. (2018). *Artificial intelligence and the 'good society': The US, EU, and UK approach*. *Science and Engineering Ethics*, 24(2), 505–528. https://doi.org/10.1007/s11948-017-9901-7

European Union Agency for Cybersecurity. (2023). *ENISA threat landscape 2023*. Publications Office of the European Union. https://www.enisa.europa.eu/publications/enisa-threat-landscape-2023

Europol. (2023). *Facing reality? Law enforcement and the challenge of deepfakes*. Publications Office of the European Union. https://www.europol.europa.eu/publications-events/publications/facing-reality-law-enforcement-and-challenge-of-deepfakes

Fair Trials. (2024, agosto). *Artificial intelligence in public security and criminal justice systems in Latin America*. Fair Trials International. https://www.fairtrials.org/app/uploads/2024/08/Artificial-intelligence-in-public-security-and-criminal-justice-systems-in-Latin-America.pdf

Floridi, L., y Cowls, J. (2019). *A unified framework of five principles for AI in society*. *Harvard Data Science Review*, 1(1), 1-13. https://doi.org/10.1162/99608f92.8cd550d1

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., y Bengio, Y. (2014). Generative adversarial networks. ArXiv, 1-9. https://doi.org/10.48550/arXiv.1406.2661

INTERPOL. (2020). *Artificial intelligence and policing: Global trends and future perspectives*. INTERPOL Innovation Centre. https://www.interpol.int/en/content/download/15586/file/INTERPOL_AI_Report_2020.pdf

Mitchell, M. (2021). *Artificial intelligence: A guide for thinking humans*. Penguin Books.

Organisation for Economic Co-operation and Development. (2022). *OECD framework for the classification of AI systems: A tool for effective AI policies*. OECD Publishing. https://doi.org/10.1787/cb6d9eca-en

Prosegur. (2024, 15 de enero). *Cybersecurity in Latin America: Legislative challenge and trends*. Prosegur. https://www.prosegur.com/en/innovation/ciber/cybersecurity-latin-america-legislative-challenge

Rishi-Bommasani, D. A., et al. (2022). On the opportunities and risks of foundation models. arXiv. https://doi.org/10.48550/arXiv.2108.07258

Reuters. (2024, 9 de octubre). *Mexico faces over half Latin American cybercrimes due largely to U.S. ties*. Reuters. https://www.reuters.com/world/americas/mexico-faces-over-half-latin-american-cybercrimes-due-largely-us-ties-2024-10-09

Russell, S. J., y Norvig, P. (2021). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson. https://www.amazon.com/Artificial-Intelligence-A-Modern-Approach/dp/0134610997

Schneier, B. (2012). *Liars and outliers: Enabling the trust that society needs to thrive*. Wiley.

Schneier, B. (2019, 12 de junio). AI can thrive in open societies. *Schneier on Security*. https://www.schneier.com/essays/archives/2019/06/ai_can_thrive_in_ope.html

Schneier, B. (2023, octubre). AI risks. *Schneier on Security*. https://www.schneier.com/blog/archives/2023/10/ai-risks.html

Schneier, B. (2023, diciembre). AI and trust. *Schneier on Security*. https://www.schneier.com/blog/archives/2023/12/ai-and-trust.html

Schneier, B. (2020). *Click here to kill everybody: Security and survival in a hyper-connected world*. W. W. Norton & Company.

Taddeo, M., y Floridi, L. (2018). How AI can be a force for good. *Science, 361*(6404), 751–752. https://doi.org/10.1126/science.aat5991

ThreatMark. (2023, 10 de diciembre). *Spain cybercrime lessons: How fraudsters exploit digital identity*. ThreatMark. https://www.threatmark.com/spain-cybercrime-lessons

ThreatMark. (2024, 7 de marzo). *Digital fraud in Latin America: Alarming trends in payments and identity theft*. ThreatMark. https://www.threatmark.com/latam-digital-fraud

United Nations Office on Drugs and Crime. (2021). *Artificial intelligence and criminal justice: Opportunities, challenges and implications for crime prevention and criminal justice*. United Nations. https://www.unodc.org/unodc/en/frontpage/2021/June/artificial-intelligence-and-criminal-justice.html

World Economic Forum. (2025). *Global cybersecurity outlook 2025*. WEFORUM. https://www.weforum.org/publications/global-cybersecurity-outlook-2025