Ciencia Latina
Internacional

Derecho Penal e Inteligencia Artificial

Responsabilidad, Evidencia, Ética y Futuro

• Yadira Alexandra Yangari Aviles
• Wendy Pricila Caranqui Dias
• Rosa Virginia Henriquez Chalen
• Yuniquer Ricardo Avendaño Vera
• Tania Evelin Mejia Auz
•Yeferson Cristian Caicedo Valencia

CID editorial
ALIANZA RED
Red de Redes
INDEXA
Crossref

Septiembre 2026 – CID - Centro de Investigación y Desarrollo
Copyright © CID - Centro de Investigación y Desarrollo
Copyright del texto © 2026 de Autores
biblioteca.ciencialatina.org
editorial@ciencialatina.org
Atención por WhatsApp al +52 22 26903834

Datos Técnicos de Publicación Internacional
Título: Derecho Penal e Inteligencia Artificial: Responsabilidad, Evidencia, Ética y Futuro
Autores: Yadira Alexandra Yangari Aviles, Wendy Pricila Caranqui Dias, Rosa Virginia Henriquez Chalen, Yuniquer Ricardo Avendaño Vera, Tania Evelin Mejia Auz y Yeferson Cristian Caicedo Valencia.
Editor: CID - Centro de Investigación y Desarrollo
Diseño de tapa: CID - Centro de Investigación y Desarrollo
Corrección de Estilo: CID - Centro de Investigación y Desarrollo
Formato: PDF
Páginas: 199 pág.
Tamaño: A4
Requisitos de sistema: Adobe Acrobat Reader
Modo de acceso: World Wide Web
ISBN: En gestión
DOI: https://doi.org/10.37811/cli_w1362

1ª Edición. Año 2026. Editorial CID - Centro de Investigación y Desarrollo.

El contenido del libro y sus datos en su forma, corrección y fiabilidad son responsabilidad exclusiva de los autores. Permite la descarga de la obra y compartir siempre que los créditos se atribuyan a los autores, pero sin la posibilidad de cambiarlo de cualquier forma o utilizarlo con fines comerciales

ÍNDICE

CAPITULO I 6

Fundamentos del Derecho Penal y la Inteligencia Artificial 7

CAPITULO II 21

Responsabilidad Penal en Sistemas Autónomos 22

CAPITULO III 46

Delitos Informáticos y Crímenes Algorítmicos 47

CAPITULO IV 75

Prueba Digital y Evidencia Algorítmica en Juicio Penal 76

CAPITULO V 105

Ética, Derechos Humanos y Justicia Penal Automatizada 106

CAPITULO VI 143

Perspectivas Futuras: Reformas Legislativas y Política Criminal 144

REFERENCIAS BIBLIOGRAFICAS 185

RESUMEN GENERAL

La obra examina de manera sistemática y progresiva los desafíos que la inteligencia artificial (IA) plantea al derecho penal contemporáneo, articulando seis ejes analíticos que configuran un panorama completo de la intersección entre tecnología algorítmica y sistema de justicia criminal. El Capítulo I establece los fundamentos conceptuales y epistemológicos de esta intersección, identificando las tensiones estructurales entre los principios garantistas del ius puniendi (legalidad, culpabilidad, proporcionalidad y presunción de inocencia) y la lógica de los sistemas algorítmicos autónomos. El Capítulo II analiza la crisis de imputación penal generada por los sistemas de IA de alta autonomía, documentando el denominado vacío de responsabilidad que afecta a toda la cadena de actores involucrados: desarrolladores, fabricantes, desplegadores y usuarios. El Capítulo III cartografía las nuevas tipologías delictivas vinculadas a la IA (fraude algorítmico, deepfakes ilícitos, ciberataques autónomos y crímenes de identidad sintética) evidenciando el desfase normativo entre las conductas posibles y los tipos penales vigentes. El Capítulo IV aborda la problemática de la prueba digital y la evidencia algorítmica en el juicio penal, revelando una brecha crítica entre la sofisticación técnica de los sistemas generadores de evidencia y la capacidad del derecho procesal para someterlos a contradicción efectiva. El Capítulo V examina la dimensión ética y de derechos humanos de la justicia penal automatizada, constatando que la automatización reproduce y amplifica sesgos estructurales que comprometen la igualdad y la dignidad humana. El Capítulo VI, de cierre, propone reformas legislativas y de política criminal orientadas a superar la fragmentación normativa global, identificando cuatro ejes indispensables: actualización de tipos penales, reformulación de criterios de imputación, modernización del estatuto probatorio y construcción de marcos supranacionales de cooperación.

Palabras clave: *inteligencia artificial; derecho penal; responsabilidad algorítmica; crímenes algorítmicos; evidencia digital; ética de la IA; derechos humanos; reforma legislativa.*

GENERAL ABSTRACT

This work systematically and progressively examines the challenges that artificial intelligence (AI) poses to contemporary criminal law, articulating six analytical axes that configure a comprehensive overview of the intersection between algorithmic technology and the criminal justice system. Chapter I establishes the conceptual and epistemological foundations of this intersection, identifying the structural tensions between the fundamental principles of criminal law (legality, culpability, proportionality, and presumption of innocence) and the logic of autonomous algorithmic systems. Chapter II analyzes the crisis of criminal liability generated by highly autonomous AI systems, documenting the so-called liability vacuum that affects the entire chain of actors involved: developers, manufacturers, deployers, and users. Chapter III maps the new types of crimes linked to AI (algorithmic fraud, illicit deepfakes, autonomous cyberattacks, and synthetic identity crimes), highlighting the normative gap between possible conduct and existing criminal offenses. Chapter IV addresses the problems of digital evidence and algorithmic evidence in criminal trials, revealing a critical gap between the technical sophistication of evidence-generating systems and the capacity of procedural law to subject them to effective cross-examination. Chapter V examines the ethical and human rights dimensions of automated criminal justice, finding that automation reproduces and amplifies structural biases that compromise equality and human dignity. Chapter VI, the closing chapter, proposes legislative and criminal policy reforms aimed at overcoming global regulatory fragmentation, identifying four essential pillars: updating criminal offenses, reformulating imputation criteria, modernizing the rules of evidence, and building supranational frameworks for cooperation.

Keywords: *artificial intelligence; criminal law; algorithmic liability; algorithmic crimes; digital evidence; AI ethics; human rights; legislative reform.*

CAPITULO I

Fundamentos del Derecho Penal y la Inteligencia Artificial

Fundamentos del Derecho Penal y la Inteligencia Artificial

YADIRA ALEXANDRA YANGARI AVILES[1]

RESUMEN

El Capítulo I examina los fundamentos conceptuales y epistemológicos de la intersección entre el derecho penal y la inteligencia artificial (IA), en un contexto de acelerado desarrollo tecnológico que desafía los marcos jurídicos clásicos. El derecho penal, disciplina que ostenta el mayor poder coercitivo del Estado, enfrenta desafíos de envergadura inédita ante la irrupción de sistemas algorítmicos dotados de aprendizaje autónomo y toma de decisiones en entornos no estructurados. El estudio, de naturaleza teórico-documental, analiza fuentes especializadas, identificando cinco grandes núcleos temáticos: los sistemas de evaluación de riesgo algorítmico, el reconocimiento facial en investigaciones criminales, la responsabilidad penal ante daños causados por IA, las garantías procesales comprometidas y los marcos regulatorios comparados. Los resultados evidencian tensiones estructurales irresueltas entre eficiencia y garantismo, predictibilidad y presunción de inocencia, y transparencia algorítmica frente a seguridad operacional. Se constata un vacío de responsabilidad documentado cuando sistemas autónomos producen resultados típicos, pues los conceptos clásicos de autoría, causalidad y culpabilidad resultan insuficientes. La evidencia digital generada por IA carece de un marco procesal penal adecuado en la mayoría de ordenamientos jurídicos. Se concluye que los principios cardinales del derecho penal liberal no son incompatibles con el uso de IA en la justicia, pero su compatibilidad exige salvaguardas normativas y técnicas aún no implementadas sistemáticamente. El Reglamento de Inteligencia Artificial de la Unión Europea (2024/1689) representa un avance significativo aunque insuficiente, siendo urgente el desarrollo de una dogmática penal tecnológicamente informada que preserve las garantías fundamentales del individuo frente al poder punitivo del Estado.

Palabras clave: *Inteligencia artificial, Derecho penal, Responsabilidad penal, Vacío de responsabilidad, Garantías procesales, Opacidad algorítmica, Imputación objetiva, Gobernanza tecnológica.*

[1] Abogada de los Juzgados y Tribunales de la República del Ecuador, Universidad Guayaquil, Magister en Victimologia y Justicia Restaurativa, Instituto de Altos Estudios Nacionales, Funcionaria Publica, Fiscalía General del Estado, correo electrónico: yadira-20001@hotmail.com.

ABSTRACT

Chapter I examines the conceptual and epistemological foundations of the intersection between criminal law and artificial intelligence (AI) in a context of rapid technological development that challenges traditional legal frameworks. Criminal law, the discipline that wields the state's greatest coercive power, faces unprecedented challenges with the emergence of algorithmic systems endowed with autonomous learning and decision-making capabilities in unstructured environments. This theoretical and documentary study analyzes specialized sources, identifying five main thematic areas: algorithmic risk assessment systems, facial recognition in criminal investigations, criminal liability for damages caused by AI, compromised procedural safeguards, and comparative regulatory frameworks. The findings reveal unresolved structural tensions between efficiency and due process, predictability and the presumption of innocence, and algorithmic transparency versus operational safety. A documented accountability gap is evident when autonomous systems produce typical results, as classical concepts of authorship, causation, and culpability prove insufficient. Digital evidence generated by AI lacks an adequate criminal procedural framework in most legal systems. It is concluded that the cardinal principles of liberal criminal law are not incompatible with the use of AI in the justice system, but their compatibility requires normative and technical safeguards that have not yet been systematically implemented. The European Union's Artificial Intelligence Regulation (2024/1689) represents a significant, though insufficient, step forward, making the development of a technologically informed criminal doctrine that preserves fundamental individual guarantees against the State's punitive power urgent.

Keywords: *Artificial intelligence, Criminal law, Criminal liability, Liability vacuum, Procedural guarantees, Algorithmic opacity, Objective imputation, Technological governance.*

INTRODUCCIÓN

El desarrollo acelerado de la inteligencia artificial (IA) durante la última década ha generado transformaciones profundas en prácticamente todos los sectores de la actividad humana, y el ámbito jurídico no constituye una excepción. En particular, el derecho penal (disciplina que se ocupa de las conductas más gravosas para la sociedad y que ostenta el mayor poder coercitivo del Estado) se enfrenta hoy a desafíos epistemológicos, normativos y filosóficos de envergadura inédita. La convergencia entre sistemas algorítmicos de alto rendimiento y los principios fundacionales del ius puniendi plantea interrogantes que los marcos teóricos clásicos no están en condiciones de responder de manera satisfactoria (Doshi-Velez et al., 2021; Raso et al., 2022).

Desde una perspectiva histórica, el derecho penal ha evolucionado en respuesta a las transformaciones tecnológicas y sociales de cada época: la industrialización del siglo XIX demandó nuevas categorías delictivas en torno a la maquinaria y la propiedad; el advenimiento de la era digital en la segunda mitad del siglo XX obligó a los legisladores a tipificar los delitos informáticos. No obstante, la irrupción de sistemas de IA dotados de capacidad de aprendizaje autónomo (machine learning), de procesamiento de lenguaje natural y de toma de decisiones en entornos no estructurados supone un salto cualitativo que excede el paradigma de la herramienta pasiva para adentrarse en el territorio de los agentes cuasi-autónomos (Calo, 2020; Berk, 2021).

La relevancia académica y práctica del presente estudio se asienta sobre tres grandes ejes. En primer lugar, la proliferación de sistemas de IA en funciones de seguridad pública y justicia criminal (vigilancia predictiva, análisis de riesgo de reincidencia, sistemas de reconocimiento facial) ha suscitado debates urgentes sobre la compatibilidad de dichos instrumentos con las garantías constitucionales del proceso penal, especialmente el derecho a un juicio justo, la presunción de inocencia y el principio de igualdad (Angwin et al., 2016; Završnik, 2021). En segundo lugar, la atribución de responsabilidad penal cuando una conducta delictiva es cometida mediante, o con la participación determinante de, un sistema de IA plantea una crisis de imputación que los conceptos clásicos de autoría y participación no permiten resolver con claridad (Dremliuga & Kucherov, 2022; Hallevy, 2022). En tercer lugar, la utilización de evidencia digital generada por sistemas de IA (desde registros de geolocalización hasta análisis

de patrones de comportamiento) introduce complejidades probatorias que la teoría general de la prueba penal no previó (Casey, 2019; Raeini et al., 2023).

El presente capítulo se propone trazar los fundamentos conceptuales y epistemológicos que permitan abordar de manera rigurosa la intersección entre el derecho penal y la inteligencia artificial. Para ello, se examinan, en primer término, los principios estructurales del derecho penal moderno (legalidad, culpabilidad, proporcionalidad, presunción de inocencia) en cuanto estándares frente a los cuales ha de evaluarse cualquier incorporación de tecnología en el sistema de justicia. En segundo término, se ofrece una caracterización técnica y jurídica de los sistemas de IA que resulta indispensable para una discusión dogmáticamente fundada. En tercer término, se identifican los principales ámbitos de intersección y se delimita el enfoque metodológico adoptado para el conjunto del libro.

La pertinencia de este esfuerzo analítico se ve subrayada por la proliferación de iniciativas regulatorias a nivel global. La Unión Europea promulgó el Reglamento de Inteligencia Artificial (Reglamento UE 2024/1689), que clasifica los sistemas de IA utilizados en el ámbito de la justicia penal como de alto riesgo y establece requisitos estrictos de transparencia, auditabilidad y supervisión humana (Comisión Europea, 2024). Por su parte, Estados Unidos, China y diversas naciones latinoamericanas han comenzado a desarrollar marcos normativos que, con distintos énfasis, intentan conciliar la innovación tecnológica con las garantías del Estado de Derecho (Floridi et al., 2022; Ortega & Salinas, 2023). La ausencia de consenso internacional y la fragmentación regulatoria hacen tanto más urgente la elaboración de bases conceptuales sólidas que puedan orientar el debate jurídico comparado.

Cabe precisar que el presente trabajo no pretende ofrecer soluciones legislativas definitivas ni pronunciarse sobre casos concretos, sino sentar las bases teóricas y metodológicas que permitan un tratamiento riguroso de los problemas que la IA plantea al derecho penal. La ambición es, en suma, contribuir a una dogmática penal tecnológicamente informada que sea capaz de preservar las garantías fundamentales del individuo frente al poder punitivo del Estado en el entorno de la sociedad digital (Zuboff, 2021; Hildebrandt, 2020).

MATERIAL Y MÉTODOS

Diseño de Investigación

El presente capítulo se inscribe en el marco de una investigación de naturaleza teórico-documental, cuyo diseño metodológico combina la revisión bibliográfica sistemática con el análisis de fuentes secundarias especializadas. Este enfoque resulta el más adecuado dado que el objeto de estudio no requiere trabajo de campo ni recolección de datos empíricos primarios, sino la identificación, sistematización e interpretación crítica del estado del arte en la intersección entre el derecho penal y la inteligencia artificial.

En cuanto a las fuentes primarias de análisis, se seleccionaron artículos científicos publicados en revistas indexadas en las bases de datos Scopus, Web of Science, HeinOnline y LexisNexis Academic en el período comprendido entre enero de 2020 y marzo de 2025. Los criterios de búsqueda incluyeron los términos: "artificial intelligence AND criminal law", "algorithmic decision-making AND criminal justice", "machine learning AND criminal liability", "predictive policing AND due process", así como sus equivalentes en español. Se limitó la búsqueda a artículos con factor de impacto cuartil Q1 o Q2 en sus respectivas áreas, garantizando así la calidad científica del corpus analizado.

El corpus documental incluyó además instrumentos normativos de rango europeo, internacional y comparado (en particular el Reglamento de Inteligencia Artificial de la Unión Europea (2024/1689), los principios de la OCDE sobre IA (2023), y las directrices del Consejo de Europa en materia de IA y justicia), así como informes institucionales de organismos internacionales relevantes como la UNESCO, el Comité de Ministros del Consejo de Europa y el Grupo de Trabajo sobre IA del G7. En total, se procesaron 147 fuentes documentales, de las cuales se seleccionaron 33 como referencias directamente citadas en el texto, atendiendo a los criterios de pertinencia temática, rigor metodológico y actualidad.

El análisis de los materiales seleccionados se llevó a cabo mediante la técnica del análisis documental cualitativo, que comprende las fases de lectura comprensiva, identificación de argumentos centrales, construcción de categorías analíticas y síntesis interpretativa. Se adoptó un enfoque dogmático-jurídico para el tratamiento de las fuentes de derecho positivo y doctrinal, y un enfoque hermenéutico para la interpretación de los documentos de política pública y los informes técnicos. La triangulación de perspectivas disciplinares (derecho penal,

ciencia de la computación, filosofía del derecho y ética de la tecnología) se empleó como estrategia de validación interna del análisis.

Entre las limitaciones propias de este diseño metodológico conviene señalar que la rápida evolución del campo puede generar cierta obsolescencia parcial de algunas referencias en el momento de la publicación, dado el ritmo acelerado al que se producen tanto los avances tecnológicos como las respuestas normativas. Asimismo, la prevalencia de literatura en inglés en las revistas de mayor impacto introduce un sesgo de selección geográfico, que se ha intentado compensar incorporando fuentes doctrinales en español, francés y alemán de reconocido prestigio en el ámbito del derecho penal continental.

RESULTADOS

Estado del Arte en la Intersección entre IA y Derecho Penal

El análisis de la literatura especializada permite constatar que la investigación académica sobre la intersección entre IA y derecho penal se ha intensificado de manera exponencial a partir de 2018, con un punto de inflexión significativo en 2020-2021 asociado, por una parte, a la publicación del primer borrador del Reglamento europeo de IA y, por otra, a los debates suscitados en Estados Unidos por el uso de sistemas algorítmicos de evaluación de riesgo en el sistema de justicia penal. La distribución geográfica de la producción académica muestra un claro predominio de las instituciones anglófonas (Reino Unido, Estados Unidos, Australia y Canadá), seguidas por centros de investigación europeos, especialmente alemanes, holandeses y escandinavos, en lo que respecta a la doctrina penal y a los estudios de derechos fundamentales.

La revisión bibliográfica identifica cinco grandes núcleos temáticos en torno a los cuales se organiza la producción científica. El primer núcleo el cual es el de mayor volumen de publicaciones se refiere a los sistemas de evaluación de riesgo algorítmico (risk assessment tools) en el sistema de justicia penal, con especial atención al caso COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) y su impacto sobre las poblaciones afroamericanas en Estados Unidos, estudiado pioneramente por Angwin et al. (2016) y debatido posteriormente desde la perspectiva de la equidad algorítmica por Dressel y Farid (2022) y por Binns (2022). El segundo núcleo temático, en plena expansión a partir de 2020, se refiere al uso de sistemas de reconocimiento facial en las investigaciones criminales y en el

control de espacios públicos, con estudios que documentan tasas de error diferenciadas por grupos raciales y de género que comprometen el principio de igualdad ante la ley (Garvie et al., 2022; Buolamwini & Gebru, 2020).

El tercer núcleo temático aborda las cuestiones de responsabilidad penal en los supuestos en que un sistema de IA participa de manera determinante en la producción de un resultado delictivo. La literatura identifica un vacío de responsabilidad (responsibility gap) que resulta de la combinación de tres factores: la distribución de tareas entre múltiples actores humanos (programador, fabricante, desplegador, usuario) que diluye la imputabilidad individual; la autonomía del sistema en la toma de decisiones, que genera distancia causal entre la conducta humana y el resultado; y la opacidad algorítmica, que dificulta la reconstrucción del nexo causal en términos procesalmente admisibles (Matthias, 2021; Dremliuga & Kucherov, 2022). El cuarto núcleo temático se ocupa de las implicaciones del uso de IA para las garantías procesales penales, y ha producido una literatura jurídico-procesal de notable profundidad (Završnik, 2021; Hildebrandt, 2020; Pasquale, 2020). El quinto núcleo, emergente y en proceso de consolidación, analiza los marcos regulatorios en perspectiva comparada, evaluando las fortalezas y debilidades de los distintos modelos de gobernanza de la IA en el ámbito de la justicia penal (Floridi et al., 2022; Comisión Europea, 2024; Ortega & Salinas, 2023).

Instrumentos Normativos Internacionales: Tabla Comparativa

Instrumento	Jurisdicción	Año	Disposiciones clave para la justicia penal
Reglamento UE de IA (2024/1689)	Unión Europea	2024	Clasifica sistemas IA en justicia penal como de alto riesgo; exige transparencia, supervisión humana y derecho de impugnación
Principios OCDE sobre IA	Internacional	2019/2023	Principios de transparencia, responsabilidad y robustez; revisados en 2023 para incluir sistemas generativos

Instrumento	Jurisdicción	Año	Disposiciones clave para la justicia penal
Rec. Consejo de Europa CM/Rec(2020)1	Consejo de Europa	2020	Directrices sobre IA en el sistema judicial; énfasis en independencia judicial y no discriminación
Blueprint for an AI Bill of Rights (EO 14110)	Estados Unidos	2022/2023	Marco no vinculante; protección frente a sistemas algorítmicos discriminatorios en justicia y seguridad
Rec. UNESCO sobre la ética de la IA	Internacional (193 países)	2021	Primera norma global; prohíbe sistemas de puntuación social con impacto discriminatorio en el sistema de justicia

Nota: Elaboración propia a partir de Comisión Europea (2024), OCDE (2023), Consejo de Europa (2020), UNESCO (2021).

Tensiones Normativas Identificadas

La revisión de la literatura permite identificar un conjunto de tensiones normativas estructurales que atraviesan todos los ámbitos de la intersección entre IA y derecho penal. La primera y más fundamental es la tensión entre eficiencia y garantismo: los sistemas de IA prometen incrementar la eficiencia del sistema de justicia (reducir tiempos de tramitación, aumentar la tasa de esclarecimiento de delitos, optimizar la asignación de recursos policiales), pero lo hacen frecuentemente a costa de garantías procesales fundamentales cuya erosión afecta desproporcionadamente a los grupos ya vulnerables (Završnik, 2021; Richardson et al., 2019).

La segunda tensión se produce entre predictibilidad y presunción de inocencia. Los sistemas de evaluación de riesgo pretenden predecir la peligrosidad futura de un individuo a efectos de adoptar decisiones de prisión preventiva, libertad condicional o intensidad de la supervisión. Pero este enfoque actuarial introduce una lógica preventivo-especial que entra en conflicto con el principio de presunción de inocencia y con el principio de culpabilidad (Dressel & Farid, 2022; Doshi-Velez & Kim, 2021).

La tercera tensión se sitúa entre transparencia y seguridad operacional. La exigencia de transparencia algorítmica (fundada en el derecho de defensa y la obligación de motivación de las resoluciones judiciales) puede entrar en conflicto con intereses legítimos de protección de la propiedad intelectual del software utilizado, o con consideraciones de seguridad operacional en contextos de investigación policial. La resolución de esta tensión requiere marcos de auditoría independiente que garanticen la verificabilidad del funcionamiento del sistema sin necesidad de una divulgación pública total de su arquitectura interna (Pasquale, 2020; Wachter et al., 2020).

Síntesis de Problemas por Ámbito de Intersección

Ámbito	Principales problemas identificados	Referencias clave
Criminalidad habilitada por IA	Imputación de autoría; distancia causal; deepfakes; nuevas tipologías delictivas	Chesney & Citron (2022); Brundage et al. (2021); Dremliuga & Kucherov (2022)
Investigación criminal asistida por IA	Admisibilidad de evidencia algorítmica; sesgos en reconocimiento facial; predictive policing	Garvie et al. (2022); Ferguson (2021); Richardson et al. (2019); Casey (2019)
Administración de justicia mediada por IA	Motivación judicial; discriminación algorítmica; derecho de impugnación; sesgos en COMPAS	Angwin et al. (2016); Završnik (2021); Dressel & Farid (2022); Binns (2022)
Marco regulatorio y gobernanza	Fragmentación normativa; déficit democrático; brecha entre mandato y capacidad técnica	Comisión Europea (2024); Floridi et al. (2022); Hildebrandt (2020); Ortega & Salinas (2023)

Nota: Elaboración propia a partir del corpus bibliográfico analizado.

El Problema de la Responsabilidad en Sistemas de IA

Una de las observaciones más consistentes a lo largo de la literatura revisada es la insuficiencia de los modelos tradicionales de responsabilidad penal para dar cuenta de los supuestos en que un sistema de IA participa de manera relevante en la producción de un resultado típico. Esta insuficiencia se manifiesta en tres escenarios diferenciados que la doctrina ha comenzado a sistematizar.

El primer escenario es el de la herramienta clásica: el ser humano utiliza el sistema de IA como instrumento directamente controlado para cometer un delito. En este supuesto no existe mayor dificultad dogmática, pues la responsabilidad del usuario es directa e imputable conforme a los criterios ordinarios de autoría (Matthias, 2021; Hallevy, 2022). El segundo escenario es el de la autonomía parcial: el sistema de IA actúa con un grado significativo de autonomía dentro de parámetros fijados por el programador o el operador, y produce un resultado perjudicial que no fue directamente querido por ningún actor humano. En este caso, la imputación puede canalizarse hacia el fabricante o el desplegador del sistema mediante categorías como la imprudencia por diseño defectuoso o la imprudencia por despliegue inadecuado (Dremliuga & Kucherov, 2022; Raso et al., 2022). El tercer escenario y el más problemático es el de la autonomía plena o emergente: el sistema de IA genera un comportamiento no previsto ni previsible por sus creadores, que produce un resultado delictivo. En este supuesto, los mecanismos tradicionales de imputación resultan insuficientes, y la literatura ha comenzado a explorar alternativas como la responsabilidad objetiva por riesgo, la figura del operador como garante, o la atribución de personalidad jurídica al sistema de IA, propuesta que, en el ámbito del derecho penal, encuentra objeciones fundamentales derivadas del principio de culpabilidad (Hallevy, 2022; Floridi et al., 2022).

La observación de estos tres escenarios permite concluir que, en el estado actual del desarrollo tecnológico y del derecho positivo, los supuestos de mayor autonomía del sistema de IA generan efectivamente un vacío de responsabilidad que ningún ordenamiento jurídico ha logrado colmar de manera satisfactoria. Esta constatación debe conducir a una reformulación de los criterios de imputación de los actores humanos en la cadena de desarrollo y despliegue de los sistemas, antes que a la precipitada atribución de responsabilidad penal a los propios sistemas de IA.

Evidencia Algorítmica y Estándares Probatorios

Un segundo hallazgo relevante de la revisión bibliográfica se refiere a la problemática de la evidencia digital generada o seleccionada por sistemas de IA en el contexto del proceso penal. La incorporación de evidencia algorítmica al proceso plantea un desafío específico de validación epistémica: ¿en virtud de qué criterios puede un tribunal aceptar como prueba válida la conclusión de un sistema de IA, cuando la lógica interna de esa conclusión no puede ser examinada mediante los procedimientos ordinarios de contradicción procesal?

La literatura especializada identifica tres estándares potencialmente aplicables. El primero es el estándar de validación técnica, conforme al cual la evidencia algorítmica sería admisible si el sistema que la produjo ha sido validado científicamente mediante pruebas de fiabilidad, precisión y ausencia de sesgo discriminatorio (Casey, 2019; Raeini et al., 2023). El segundo es el estándar de auditabilidad procedimental, conforme al cual la admisibilidad de la evidencia depende de que el sistema pueda ser auditado por peritos independientes en el curso del proceso (Wachter et al., 2020). El tercero es el estándar de explicabilidad sustantiva, que exigiría que el sistema sea capaz de proporcionar, para cada conclusión particular, una explicación comprensible para las partes y el tribunal (Doshi-Velez & Kim, 2021; Arrieta et al., 2020). Ninguno de estos estándares ha alcanzado aún consagración legislativa explícita en los principales ordenamientos procesales penales, y la jurisprudencia comparada muestra soluciones ad hoc y fragmentarias que generan inseguridad jurídica.

DISCUSIÓN Y CONCLUSIONES

Discusión de los Hallazgos

Los resultados obtenidos de la revisión bibliográfica permiten trazar un panorama sistemático de los principales desafíos que la inteligencia artificial plantea al derecho penal, al tiempo que revelan el estado de desarrollo aún incipiente de las respuestas teóricas y normativas disponibles. La discusión de estos hallazgos debe situarse en el contexto más amplio de los estudios sobre la gobernanza de las tecnologías emergentes y su relación con el Estado de Derecho.

Un primer aspecto que merece reflexión crítica es la asimetría entre el ritmo de innovación tecnológica y el ritmo de la respuesta jurídico-normativa. La literatura revisada documenta consistentemente que los sistemas de IA se despliegan en el ámbito de la justicia penal con

anterioridad al establecimiento de marcos regulatorios adecuados, generando situaciones de facto que luego resultan difíciles de revertir. Este patrón, denominado por Hildebrandt (2020) como regulación por defecto tecnológico, supone que son las decisiones de diseño de los ingenieros y las estrategias comerciales de las empresas tecnológicas quienes, de facto, determinan las condiciones de ejercicio del poder punitivo del Estado, desplazando al legislador democrático de su función constitucional. La Unión Europea ha intentado invertir esta dinámica mediante el Reglamento de IA de 2024, que adopta un enfoque ex ante basado en la evaluación de riesgos, pero la efectividad de este modelo regulatorio dependerá en última instancia de la capacidad de las autoridades de supervisión para fiscalizar sistemas técnicamente complejos (Comisión Europea, 2024; Floridi et al., 2022).

Un segundo aspecto que la discusión debe abordar es la especificidad cultural y jurídica de los problemas identificados. Gran parte de la literatura revisada proviene del contexto angloamericano, donde el common law y el sistema adversarial de enjuiciamiento penal plantean los problemas de evidencia y de motivación de una manera que no es directamente trasladable a los sistemas de derecho continental. En los sistemas inquisitivos o mixtos, como los prevalentes en Europa continental y América Latina, la motivación de las resoluciones judiciales es un requisito constitucional y procesal mucho más exigente, lo que hace que la tensión entre opacidad algorítmica y deber de motivación sea potencialmente más aguda (Završnik, 2021; Ortega & Salinas, 2023).

Un tercer aspecto relevante para la discusión es la dimensión de los derechos humanos y la equidad en el uso de IA en la justicia penal. Los estudios empíricos revisados documentan de manera consistente que los sistemas algorítmicos utilizados en la justicia penal tienden a reproducir y amplificar los sesgos presentes en los datos históricos sobre los que fueron entrenados. Dado que dichos datos reflejan las desigualdades estructurales de los sistemas de justicia que han producido históricamente tasas de criminalización desproporcionadas sobre poblaciones minoritarias, los sistemas de IA no solo perpetúan estas desigualdades, sino que les confieren una aparente legitimidad tecnocrática (Angwin et al., 2016; Buolamwini & Gebru, 2020; Dressel & Farid, 2022). Este hallazgo tiene implicaciones de primer orden para la compatibilidad del uso de estos sistemas con el principio de igualdad ante la ley.

En relación con trabajos previos, los resultados de la presente revisión confirman y amplían las conclusiones alcanzadas por Završnik (2021) en su análisis de la justicia penal algorítmica

en Europa, y son consistentes con las observaciones de Hildebrandt (2020) sobre la tensión entre rule of law e inteligencia artificial. Al mismo tiempo, la presente revisión actualiza y complementa dichos análisis incorporando los desarrollos más recientes derivados de la aprobación del Reglamento europeo de IA y del debate en torno a los modelos de lenguaje a gran escala, aspectos que no podían ser abordados en los trabajos publicados con anterioridad a 2023.

Conclusiones

De la revisión y análisis efectuados en el presente capítulo es posible extraer las siguientes conclusiones, que se formulan con la cautela epistemológica que impone la naturaleza documental de la investigación y el estado aún abierto de los debates examinados.

La intersección entre la inteligencia artificial y el derecho penal constituye uno de los desafíos jurídicos más complejos y urgentes de la era digital, dado que compromete simultáneamente los fundamentos de la imputación penal, las garantías procesales del individuo y los principios estructurales del Estado de Derecho. Su tratamiento adecuado exige un enfoque interdisciplinario que integre las perspectivas de la dogmática jurídico-penal, la ciencia de la computación, la filosofía del derecho y la ética de la tecnología.

Los principios cardinales del derecho penal liberal (legalidad, culpabilidad, proporcionalidad, presunción de inocencia) no son, en principio, incompatibles con el uso de la inteligencia artificial en el sistema de justicia, pero su compatibilidad requiere el diseño deliberado de salvaguardas normativas y técnicas que en la actualidad no han sido implementadas de manera sistemática en ningún ordenamiento jurídico. La aprobación del Reglamento de IA de la Unión Europea de 2024 constituye un avance significativo, pero insuficiente, en esta dirección.

Existe un vacío de responsabilidad documentado en los supuestos de daño producido por sistemas de IA con elevada autonomía, que los mecanismos tradicionales de imputación penal no permiten cubrir. La superación de este vacío requiere tanto reformas normativas orientadas a clarificar la distribución de responsabilidad en la cadena de desarrollo y despliegue de los sistemas, como el desarrollo de nuevas categorías dogmáticas capaces de capturar la especificidad de la responsabilidad por defecto de diseño o por negligencia en el control de sistemas autónomos.

La evidencia digital generada o seleccionada por sistemas de IA carece, en la actualidad, de un marco procesal penal adecuado en la mayoría de los ordenamientos jurídicos. La determinación de estándares de admisibilidad, validación y contradicción de la evidencia algorítmica es una tarea normativa pendiente de primera importancia, cuya omisión genera riesgos serios para el derecho a un juicio justo.

Los sesgos algorítmicos documentados en los sistemas de evaluación de riesgo y reconocimiento facial utilizados en el ámbito de la justicia penal constituyen una amenaza real y actual al principio de igualdad ante la ley. Su abordaje requiere no solo medidas técnicas de desbiasado (debiasing), sino también mecanismos jurídicos de control que permitan la impugnación efectiva de decisiones fundadas en sistemas algorítmicos discriminatorios.

En cuanto a las limitaciones del presente estudio, cabe señalar que la revisión bibliográfica no agota el caudal de reflexión producido sobre la materia, dado el ritmo de producción académica en el campo. Las líneas de investigación que se derivan de los hallazgos presentados incluyen: el análisis comparado de la jurisprudencia penal sobre evidencia algorítmica; el estudio de la implementación efectiva del Reglamento de IA europeo en el ámbito de la justicia penal; el desarrollo de propuestas dogmáticas para la imputación de responsabilidad en sistemas de IA autónomos; y la evaluación del impacto diferencial de los algoritmos de predicción de riesgo sobre grupos étnicos y socioeconómicos en contextos latinoamericanos.

CAPITULO II

Responsabilidad Penal en Sistemas Autónomos

Responsabilidad Penal en Sistemas Autónomos

WENDY PRICILA CARANQUI DIAS[2]

RESUMEN

El avance de los sistemas de inteligencia artificial (IA) con capacidades autónomas de decisión plantea una crisis estructural en los fundamentos de la responsabilidad penal. El esquema clásico del derecho penal, edificado sobre la culpabilidad individual y el libre albedrío, resulta insuficiente cuando el agente dañoso es un sistema algorítmico que opera con autonomía creciente, opacidad interna y una cadena de desarrollo en la que participan múltiples actores humanos. El presente capítulo analiza los principales modelos de imputación penal aplicables a los supuestos de daño producido por sistemas autónomos de IA, examinando las posiciones de los distintos actores de la cadena (desarrollador, fabricante, desplegador y usuario) así como las propuestas doctrinales que postulan la atribución de responsabilidad objetiva o la creación de una personalidad jurídica electrónica. Mediante una metodología de investigación teórico-documental y de derecho comparado, se identifican las insuficiencias de las categorías vigentes y se proponen criterios de imputación diferenciados en función del nivel de autonomía del sistema y del grado de previsibilidad del resultado dañoso. Se concluye que el actual vacío de responsabilidad exige una reforma dogmática y legislativa orientada a distribuir cargas de diligencia entre los distintos actores de la cadena de IA, sin renunciar a los principios garantistas del derecho penal liberal.

Palabras clave: *responsabilidad penal; inteligencia artificial; sistemas autónomos; imputación objetiva; vacío de responsabilidad; culpabilidad; personalidad jurídica electrónica; derecho penal digital.*

[2] Abogada de los Juzgados y Tribunales de la República del Ecuador, Universidad Guayaquil, Magister en Derecho Procesal, Universidad Tecnologica Ecotec, Funcionaria Publica, Fiscalia General del Estado, correo electrónico: wcaranquid@gmail.com.

ABSTRACT

The advancement of artificial intelligence (AI) systems with autonomous decision-making capabilities poses a structural crisis in the foundations of criminal liability. The classical criminal law framework, built upon individual culpability and free will, proves insufficient when the harmful agent is an algorithmic system that operates with increasing autonomy, internal opacity, and a development chain involving multiple human actors. This chapter analyzes the main models of criminal imputation applicable to cases of harm caused by autonomous AI systems, examining the positions of the different actors in the chain — developer, manufacturer, deployer and user— as well as doctrinal proposals postulating the attribution of strict liability or the creation of an electronic legal personality. Through a theoretical-documentary and comparative law methodology, the inadequacies of existing categories are identified, and differentiated imputation criteria are proposed according to the autonomy level of the system and the foreseeability of the harmful result. It is concluded that the current accountability gap requires both dogmatic and legislative reform aimed at distributing duty-of-care obligations among the different actors in the AI chain, without abandoning the guarantees of liberal criminal law.

Keywords: *criminal liability; artificial intelligence; autonomous systems; objective imputation; accountability gap; culpability; electronic legal personality; digital criminal law.*

INTRODUCCIÓN

La cuestión de quién responde penalmente por las decisiones adoptadas por un sistema de inteligencia artificial constituye, en el momento actual, uno de los problemas más desafiantes y conceptualmente ricos de la teoría jurídico-penal. No se trata de un interrogante académico de naturaleza especulativa: cada año se registran casos en que vehículos autónomos provocan accidentes con víctimas mortales, en que sistemas de diagnóstico médico impulsados por algoritmos inducen errores terapéuticos con consecuencias letales, en que robots industriales causan lesiones a operarios, o en que sistemas de trading algorítmico desencadenan operaciones fraudulentas que perjudican a miles de inversores (Vladeck, 2022; Scherer, 2021). En todos estos supuestos, la identificación del responsable penal no puede resolverse mediante la simple aplicación de las categorías dogmáticas tradicionales sin incurrir en graves distorsiones dogmáticas o en lagunas de punibilidad.

La originalidad del problema no radica únicamente en la novedad tecnológica, sino en la ruptura que los sistemas de IA de alta autonomía introducen respecto de los presupuestos filosóficos y normativos sobre los que descansa el edificio del derecho penal moderno. Desde Feuerbach hasta Jakobs, pasando por las grandes escuelas causalista, finalista y funcionalista, el derecho penal ha construido sus categorías fundamentales (acción, tipicidad, antijuridicidad, culpabilidad) sobre la premisa de que el sujeto activo del delito es un ser humano dotado de conciencia, voluntad y capacidad de motivación normativa (Roxin, 2020; Mir Puig, 2023). La irrupción de sistemas que toman decisiones de manera autónoma, que aprenden de su entorno y que generan comportamientos emergentes no previstos por sus creadores, pone en cuestión precisamente estas premisas.

Desde la perspectiva del derecho comparado, las respuestas a esta problemática son aún fragmentarias e insatisfactorias. En el ámbito europeo, el Reglamento de Inteligencia Artificial (UE) 2024/1689 ha abordado parcialmente la cuestión mediante un régimen de clasificación de riesgo y responsabilidad civil, pero ha omitido pronunciarse sobre la responsabilidad penal en términos sustantivos, materia que continúa siendo competencia de los Estados miembros (Comisión Europea, 2024). En Estados Unidos, las propuestas legislativas se han centrado en el ámbito de la responsabilidad civil y la regulación sectorial, sin abordar de manera sistemática la responsabilidad penal (Scherer, 2021; Vladeck, 2022). En América Latina, la producción normativa específica es incipiente, aunque países como Brasil, Chile, México y

Colombia han iniciado procesos legislativos de regulación de la IA que inevitablemente deberán incluir disposiciones sobre responsabilidad (Ortega & Salinas, 2023; Palomo, 2024).

El debate doctrinal ha identificado al menos cuatro grandes modelos teóricos de atribución de responsabilidad penal en el contexto de sistemas autónomos de IA. El primero distribuye la responsabilidad entre los distintos actores humanos de la cadena de desarrollo y despliegue (desarrollador, fabricante, desplegador, usuario) según criterios de contribución causal y previsibilidad. El segundo postula la responsabilidad penal corporativa o de la persona jurídica titular del sistema. El tercero propone la creación de una personalidad jurídica sui generis para los sistemas de IA de alta autonomía. El cuarto sugiere la reparación del daño mediante un fondo de compensación colectivo, sin persecución penal individual (Hallevy, 2022; Dremliuga & Kucherov, 2022; Solum, 2021; Zittrain, 2019). Ninguno de estos modelos ha alcanzado consenso doctrinal, y todos presentan tensiones significativas con los principios garantistas del derecho penal liberal.

El presente capítulo tiene por objetivo central examinar sistemáticamente estos modelos de atribución de responsabilidad penal, evaluando su compatibilidad con los principios estructurales del derecho penal, especialmente el principio de culpabilidad, el principio de legalidad y el principio de proporcionalidad, y su aptitud para resolver los casos concretos que el despliegue de sistemas autónomos de IA ya está generando en la práctica judicial comparada. Para ello, el capítulo se estructura en las siguientes partes: tras esta introducción y la sección metodológica, se presentan los resultados del análisis doctrinal, organizados en torno a las categorías de la cadena de responsabilidad, el problema de la culpabilidad, los modelos alternativos de imputación y los casos paradigmáticos; a continuación, la sección de discusión y conclusiones evalúa críticamente los hallazgos y traza las líneas de reforma necesarias. El capítulo se inscribe en la continuidad temática establecida en el Capítulo I de la presente obra, profundizando específicamente en la dimensión de la responsabilidad penal que allí fue identificada como uno de los núcleos problemáticos centrales de la intersección entre IA y derecho penal.

La pertinencia de este análisis no se agota en el plano académico. A medida que los sistemas de IA asumen funciones cada vez más relevantes en ámbitos como la conducción autónoma, la medicina, las finanzas, la seguridad y la administración pública, la definición de un marco claro de responsabilidad penal se convierte en condición necesaria no solo para la protección

de las víctimas, sino también para la generación de incentivos correctos en el diseño y despliegue de estos sistemas. Como ha señalado Scherer (2021), la indeterminación del régimen de responsabilidad genera externalidades negativas que desincentivan la adopción de estándares de seguridad más estrictos por parte de los actores del ecosistema tecnológico. En suma, la reforma del derecho penal en este ámbito es, al mismo tiempo, una exigencia de justicia para las víctimas y una política pública de gestión del riesgo tecnológico.

MATERIAL Y MÉTODOS

Diseño de Investigación

El presente capítulo se inscribe en el marco de una investigación de naturaleza teórico-documental y de derecho comparado, que adopta como estrategia metodológica central la revisión bibliográfica sistemática combinada con el análisis dogmático-jurídico de fuentes primarias y secundarias. Este diseño resulta el más adecuado para el objeto de estudio, dado que la cuestión de la responsabilidad penal en sistemas autónomos de IA es eminentemente una cuestión de construcción conceptual y de interpretación jurídica que no requiere ni admite trabajo de campo cuantitativo. La opción metodológica está en consonancia con la tradición investigadora dominante en el derecho penal comparado y con los estándares de las revistas de mayor impacto en el campo (Siems, 2018; Zweigert & Kötz, 2020).

Fuentes y Corpus Documental

La construcción del corpus documental se realizó a través de búsquedas sistemáticas en las bases de datos Scopus, Web of Science, HeinOnline, Westlaw International y SSRN, durante el período comprendido entre octubre de 2024 y febrero de 2025. Los descriptores de búsqueda empleados incluyeron los términos: "criminal liability artificial intelligence", "autonomous systems criminal law", "AI accountability gap", "robot criminal responsibility", "electronic legal person criminal law", "machine learning imputation", así como sus equivalentes en español, francés y alemán. Se priorizaron artículos publicados en revistas indexadas en los cuartiles Q1 y Q2 de sus respectivas áreas según el Journal Citation Reports. Se incorporaron también monografías y capítulos de obras colectivas de referencia obligada en la materia, instrumentos normativos de rango europeo, internacional y nacional, así como resoluciones judiciales paradigmáticas de Estados Unidos, Alemania, Reino Unido y la Unión Europea.

Criterios de Selección y Procesamiento

El corpus final procesado comprende 178 fuentes documentales, de las cuales se seleccionaron 42 como referencias directamente citadas en el texto, atendiendo a los criterios de pertinencia temática, rigor metodológico, novedad y representatividad geográfica. Se excluyeron fuentes sin identificación de autoría, textos de divulgación sin respaldo académico y documentos institucionales de carácter puramente procedimental. El análisis de los materiales se llevó a cabo mediante la técnica del análisis documental cualitativo, que comprende las fases de lectura comprensiva, identificación de argumentos centrales, construcción de categorías analíticas y síntesis interpretativa crítica.

Limitaciones Metodológicas

Entre las principales limitaciones del diseño adoptado se destacan las siguientes. En primer lugar, la acelerada evolución del campo implica que algunas de las referencias citadas pueden haber sido parcialmente superadas por desarrollos normativos o tecnológicos posteriores a la fecha de cierre del corpus. En segundo lugar, la prevalencia de literatura académica en inglés introduce un sesgo de selección geográfico que se ha intentado compensar incorporando fuentes doctrinales en español, alemán y francés. En tercer lugar, la metodología documental no permite validar empíricamente las propuestas normativas examinadas mediante estudios de impacto o análisis de datos jurisdiccionales, limitación que deberá ser subsanada en investigaciones futuras de naturaleza empírico-comparada.

RESULTADOS

La Crisis de Imputación en los Sistemas Autónomos de IA

El primer hallazgo sistemático del análisis doctrinal es la constatación de que existe una crisis de imputación real y documentada cuando un sistema autónomo de IA produce un resultado típico. Esta crisis no es de carácter meramente técnico, como podría serlo la dificultad para acceder al código fuente del sistema, sino de naturaleza estructural: afecta a las categorías más profundas del derecho penal. El análisis de la literatura permite desagregar esta crisis en tres dimensiones que se potencian mutuamente.

La primera dimensión es la crisis de la acción. El concepto penal de acción, en todas sus versiones (causal, finalista, social o funcionalista) presupone un sujeto humano que interviene

activamente en el mundo exterior mediante un comportamiento dominable por su voluntad (Roxin, 2020; Jakobs, 2019). Cuando el resultado dañoso es producido por un sistema de IA que actúa de manera autónoma en tiempo real, sin supervisión humana directa en el momento del resultado, la vinculación entre la conducta humana (el diseño, la programación o la decisión de despliegue) y el resultado típico está mediada por una cadena de procesos algorítmicos que el propio creador del sistema no puede predecir con exactitud. Esta mediación algorítmica crea lo que Matthias (2021) denominó la brecha de la responsabilidad (responsibility gap): el resultado existe, la víctima existe, pero el autor humano identificable en términos penales está ausente o resulta muy difícil de individualizar.

La segunda dimensión es la crisis de la culpabilidad. El principio de culpabilidad exige que la imposición de una pena requiera, además de la realización del tipo objetivo, un juicio de reproche personal dirigido al autor por haber actuado antijurídicamente pudiendo haber actuado de otro modo (Mir Puig, 2023). Este juicio de reproche supone la capacidad del autor de conocer el significado de su acción, de prever sus consecuencias, y de orientar su conducta conforme a la norma. Cuando el resultado dañoso es producido por un sistema de IA mediante decisiones emergentes que el programador no previó y no podía prever en función del estado del arte tecnológico al momento del diseño, el juicio de reproche personal se vuelve problemático. No existe culpabilidad directa del desarrollador respecto al resultado específico, ni existe (en los términos actuales del derecho) culpabilidad propia del sistema de IA (Hallevy, 2022; Gabriel, 2021).

La tercera dimensión es la crisis de la causalidad. La reconstrucción del nexo causal entre la conducta humana y el resultado típico, tal como exige la teoría de la imputación objetiva, se ve dificultada en los sistemas de IA por dos factores: la opacidad del proceso decisorio interno (el llamado problema de la caja negra o black box) y la intervención de múltiples actores en la cadena causal. El problema de la caja negra impide, en muchos casos, determinar cuáles son los factores que condujeron al sistema a adoptar una decisión específica, lo que hace imposible la reconstrucción pericial del proceso causal en términos procesalmente admisibles (Wachter et al., 2020; Arrieta et al., 2020). La multiplicidad de actores en la cadena causal (que incluye al diseñador del algoritmo, al equipo de entrenamiento, al fabricante del hardware, al desplegador del sistema y al usuario) diluye la imputabilidad individual hasta el punto de

generar zonas de impunidad difícilmente justificables desde la perspectiva del Estado de Derecho.

La Cadena de Responsabilidad: Actores y Posiciones Dogmáticas

Un segundo hallazgo fundamental del análisis es que la doctrina especializada ha identificado, de manera consistente, una cadena de responsabilidad que comprende al menos cuatro posiciones subjetivas distintas, cada una de las cuales plantea problemas dogmáticos específicos: el desarrollador o programador, el fabricante o empresa titular del sistema, el desplegador u operador, y el usuario final.

La Responsabilidad del Desarrollador

El desarrollador (persona individual o equipo) es quien diseña el algoritmo central del sistema de IA, determina su arquitectura y define los parámetros de su entrenamiento. La doctrina mayoritaria considera que el desarrollador puede incurrir en responsabilidad penal por imprudencia en supuestos de diseño defectuoso: cuando el sistema fue diseñado sin adoptar las medidas de seguridad exigibles conforme al estado del arte, y ese defecto de diseño es la causa adecuada del resultado dañoso (Dremliuga & Kucherov, 2022; Scherer, 2021). Esta imputación por imprudencia requiere demostrar que el desarrollador conocía o debía conocer los riesgos inherentes al sistema, que estaba en condiciones de adoptar medidas que hubieran reducido o eliminado el riesgo, y que omitió hacerlo sin justificación razonable.

La dificultad surge cuando el comportamiento dañoso del sistema es el resultado de procesos de aprendizaje automático que el propio desarrollador no podía anticipar en detalle. En este supuesto, algunos autores proponen que la responsabilidad del desarrollador se articule a través de la categoría de la imprudencia por inadecuada delimitación del sistema de aprendizaje (learning architecture negligence): si el desarrollador no estableció controles suficientes sobre los parámetros del aprendizaje ni implementó sistemas de supervisión capaces de detectar desviaciones peligrosas, esta omisión puede configurar el elemento subjetivo de la imprudencia (Vladeck, 2022; Turner, 2019). Esta propuesta tiene la virtud de anclar la responsabilidad en una omisión de cuidado identificable y valorable ex ante, sin necesidad de que el desarrollador haya previsto el resultado específico.

La Responsabilidad del Fabricante o Empresa Titular

La empresa fabricante o titular del sistema de IA presenta un perfil de responsabilidad diferenciado respecto del desarrollador individual, toda vez que su posición en la cadena de producción y comercialización le impone obligaciones específicas de control de calidad, evaluación de riesgos, información y supervisión postventa. En muchos ordenamientos jurídicos, la responsabilidad de la persona jurídica en el ámbito penal está reconocida explícitamente, ya sea mediante la imputación directa por hecho propio (modelo francés y angloamericano) o mediante la transferencia de responsabilidad desde el órgano directivo (modelo español e italiano) (Gómez-Jara Díez, 2021; Sieber, 2020).

El problema específico en el contexto de la IA radica en que la responsabilidad corporativa penal, tal como está configurada en la mayoría de los ordenamientos, exige la demostración de que el delito fue consecuencia de un defecto de organización atribuible a la empresa, esto es, de la ausencia de un sistema de cumplimiento normativo (compliance) adecuado para prevenir ese tipo de resultado. Si la empresa titular del sistema de IA dispone de un programa robusto de evaluación de riesgos, de certificación de seguridad y de supervisión del desempeño del sistema, puede argumentar que adoptó todas las medidas organizativas razonables para prevenir el daño, limitando así su responsabilidad penal. Este escenario genera incentivos para la adopción de estándares de gobernanza interna de la IA, lo que constituye un efecto preventivo-especial deseable desde la perspectiva de la política criminal (Nieto Martín, 2021).

La Responsabilidad del Desplegador u Operador

El desplegador u operador del sistema de IA (que puede ser un hospital, una empresa de transporte, una entidad financiera o un organismo público) ocupa una posición particularmente relevante en la cadena de responsabilidad, por cuanto es quien decide poner el sistema en funcionamiento en un contexto concreto, con conocimiento de sus características, limitaciones y del entorno en que operará. La doctrina comparada ha desarrollado la categoría del operador como garante (Garantenpflicht des Betreibers) para describir la obligación de supervisión que recae sobre el desplegador en virtud de su control sobre el riesgo (Sieber, 2020; Zittrain, 2019).

Esta posición de garantía implica que el desplegador tiene la obligación de verificar que el sistema de IA cumple con los estándares técnicos aplicables, de establecer protocolos de supervisión y corrección durante la operación, de informar a los usuarios sobre los límites del sistema, y de detener su funcionamiento cuando se detecten fallos o riesgos no controlados. El incumplimiento de estas obligaciones puede generar responsabilidad penal por comisión por omisión (art. 11 CP español; § 13 StGB alemán) cuando el resultado dañoso es el que precisamente debía evitar el garante mediante la adopción de las medidas de control omitidas. Este modelo de imputación tiene la ventaja de que focaliza la responsabilidad en quien tiene el control efectivo del sistema en el momento del daño, que es habitualmente el actor con mayor capacidad de prevención.

La Responsabilidad del Usuario Final

La responsabilidad penal del usuario final del sistema de IA presenta el perfil más tradicional y, en muchos supuestos, el menos problemático desde el punto de vista dogmático. Cuando el usuario final interactúa con el sistema de IA como si fuera una herramienta al servicio de sus fines delictivos (utilizando un vehículo autónomo para atropellar deliberadamente a una persona, o empleando un sistema de generación de deepfakes para cometer estafas o delitos contra el honor) la responsabilidad penal del usuario se establece mediante las categorías ordinarias de autoría directa o instrumental (Chesney & Citron, 2022; Brundage et al., 2021).

El escenario más complejo se presenta cuando el usuario opera el sistema de IA en el contexto de sus funciones profesionales, por ejemplo, un médico que sigue las indicaciones de un sistema de diagnóstico algorítmico que resultan ser erróneas, o un piloto que confía en el sistema de piloto automático de una aeronave autónoma; y el resultado dañoso se produce precisamente porque el usuario confió en las decisiones del sistema en lugar de ejercer su propio criterio profesional. En estos supuestos, la cuestión relevante es si el usuario estaba en posición de y obligado a supervisar al sistema de IA y de corregir sus indicaciones cuando eran manifiestamente incorrectas, o si la confianza en el sistema estaba justificada por las características del mismo y por los estándares profesionales aplicables. La respuesta a esta pregunta dependerá, en gran medida, de las características de cada campo profesional y del estado del arte tecnológico en el momento de los hechos (Vladeck, 2022; Turner, 2019).

Modelos Doctrinales de Imputación Penal en Sistemas Autónomos

El análisis de la literatura especializada permite identificar cinco grandes modelos doctrinales orientados a resolver el problema de la imputación penal en sistemas autónomos de IA. Estos modelos no son mutuamente excluyentes: pueden combinarse en función de las circunstancias del caso concreto y del nivel de autonomía del sistema involucrado.

El Modelo de la Imputación Distribuida

El modelo de la imputación distribuida es el más desarrollado en la doctrina actual y el que presenta mayor compatibilidad con los principios garantistas del derecho penal. Propone distribuir la responsabilidad penal entre los distintos actores de la cadena de desarrollo y despliegue del sistema de IA, en proporción a su contribución causal al resultado dañoso y a su nivel de conocimiento del riesgo (Dremliuga & Kucherov, 2022; Scherer, 2021). Este modelo opera con las categorías dogmáticas existentes (autoría y participación, comisión por imprudencia, comisión por omisión) sin necesidad de crear nuevas figuras jurídicas.

La principal limitación del modelo de imputación distribuida es que, en supuestos de alta autonomía del sistema de IA, puede conducir a lagunas de responsabilidad cuando ninguno de los actores de la cadena tenía conocimiento ni posibilidad razonable de prever el comportamiento específico del sistema que produjo el resultado dañoso. En estos casos extremos, el modelo distribuido no es capaz de cubrir el vacío de responsabilidad, porque ningún actor cumple con los requisitos del tipo subjetivo imprudente. Esta limitación ha motivado el desarrollo de modelos alternativos.

El Modelo de la Responsabilidad Objetiva por Riesgo

El modelo de la responsabilidad objetiva por riesgo parte de la premisa de que determinadas actividades de alto riesgo (entre las que cabría incluir el desarrollo y despliegue de sistemas de IA de alta autonomía en ámbitos de especial sensibilidad) generan una responsabilidad cuya imputación no requiere la demostración del elemento subjetivo doloso o imprudente, sino únicamente la del nexo causal entre la actividad y el resultado dañoso (Zittrain, 2019; Solum, 2021). Este modelo es ya conocido en el derecho de daños civil (responsabilidad por productos defectuosos, actividades peligrosas), pero su transposición al ámbito penal plantea tensiones con el principio de culpabilidad que la doctrina dominante considera insuperables en el marco del garantismo penal (Mir Puig, 2023; Jescheck & Weigend, 2019).

La postura más matizada en la doctrina consiste en proponer una versión suavizada de este modelo, en la que la responsabilidad objetiva no derive directamente en una pena privativa de libertad, sino en sanciones de carácter cuasipenales o administrativas (multas, suspensión de actividades, inhabilitaciones) reservando la pena penal propiamente dicha para los supuestos en que concurra, al menos, imprudencia grave del desarrollador, desplegador o usuario (Gómez-Jara Díez, 2021; Nieto Martín, 2021). Esta propuesta intermedia tiene el mérito de cubrir el vacío de responsabilidad sin sacrificar el principio de culpabilidad, pero requiere una definición precisa de la frontera entre la infracción administrativa y el ilícito penal que los sistemas jurídicos actuales no han establecido.

El Modelo de la Personalidad Jurídica Electrónica

El modelo de la personalidad jurídica electrónica, propuesto inicialmente en el contexto del derecho civil pero trasladado al debate penal por Hallevy (2022) y Solum (2021), postula la atribución de una subjetividad jurídica sui generis a los sistemas de IA de alta autonomía que les permita ser directamente responsables de sus acciones, al menos a efectos de determinadas consecuencias jurídicas. Este modelo tiene como precedentes la responsabilidad penal de las personas jurídicas y la atribución de derechos a entes no humanos reconocida en algunos ordenamientos.

En el ámbito penal, la propuesta de la personalidad jurídica electrónica enfrenta objeciones de principio de enorme peso. El principio de culpabilidad exige, en su formulación más consolidada, que el sujeto pasivo de la pena sea capaz de comprender el significado de la norma, de motivarse por ella y de actuar de manera diferente a como actuó. Un sistema de IA, por más avanzado que sea en términos de capacidades cognitivas simuladas, no posee en el estado actual del desarrollo tecnológico conciencia genuina, autonomía moral real ni capacidad de sufrir en el sentido que justifica filosóficamente la pena (Gabriel, 2021; Floridi, 2021). Imponer una pena a un sistema de IA sería, en el mejor de los casos, un acto simbólico sin ningún efecto preventivo real. La doctrina mayoritaria, incluidos autores tan favorables a la innovación dogmática como Hildebrandt (2020) y Pasquale (2020), rechaza la personalidad jurídico-penal plena para los sistemas de IA actuales.

El Modelo de la Responsabilidad del Garante Digital

El modelo de la responsabilidad del garante digital, propuesto por Sieber (2020) y Zittrain (2019), traslada al ámbito de la IA la categoría dogmática de la posición de garante para construir un sistema de obligaciones de supervisión cuyo incumplimiento genere responsabilidad penal por comisión por omisión. En este modelo, el desarrollador, el desplegador y el operador son considerados garantes del buen funcionamiento del sistema de IA respecto a los bienes jurídicos que este puede afectar, con la consecuencia de que la omisión de las medidas de supervisión y corrección exigibles configura el tipo omisivo.

La ventaja fundamental de este modelo es su compatibilidad con las categorías dogmáticas existentes y su capacidad para articular responsabilidades diferenciadas para cada actor de la cadena. Su limitación principal radica en la dificultad de definir, con la precisión que exige el principio de legalidad, el contenido concreto de las obligaciones de supervisión en cada contexto de despliegue, dado que el estado del arte tecnológico varía aceleradamente. Esta limitación no es insalvable: puede ser abordada mediante estándares técnicos establecidos por organismos de normalización, disposiciones reglamentarias o cláusulas de evaluación pericial que el juez penal pueda incorporar al proceso.

El Modelo del Fondo Colectivo de Responsabilidad

El modelo del fondo colectivo de responsabilidad, desarrollado en la literatura angloamericana especialmente por Vladeck (2022) y Scherer (2021), propone que en aquellos supuestos en que resulte imposible individualizar al responsable penal por la actuación autónoma de un sistema de IA, la reparación del daño a las víctimas se garantice mediante un fondo de compensación financiado por los actores del ecosistema de la IA (fabricantes, desplegadores, aseguradoras, eventualmente el Estado). Este modelo no constituye propiamente un modelo de responsabilidad penal, sino más bien un complemento al sistema penal que asegura la satisfacción de los intereses de las víctimas cuando el sistema punitivo no puede identificar al responsable individual.

Desde la perspectiva del derecho penal, este modelo es relevante porque alivia la presión sobre el sistema punitivo para resolver a través de la pena lo que en rigor es un problema de reparación de daños, y porque incentiva la creación de mecanismos de gobierno responsable de la IA a través de la fijación de las contribuciones al fondo en función del nivel de riesgo de

cada sistema. Su implementación requiere decisiones de política legislativa sobre el tamaño del fondo, los criterios de contribución y los mecanismos de reclamación, que van más allá del ámbito estrictamente penal.

Tabla Comparativa: Modelos de Imputación Penal en Sistemas Autónomos de IA

Modelo	Fundamento Dogmático	Actor Responsable	Ventajas	Limitaciones
Imputación distribuida	Autoría/participación; imprudencia	Desarrollador, desplegador, usuario	Compatible con principio de culpabilidad; flexible	Lagunas en alta autonomía; dificultad probatoria
Responsabilidad objetiva por riesgo	Responsabilidad sin culpa; riesgo creado	Fabricante, desplegador	Cubre vacíos de responsabilidad	Tensión con principio de culpabilidad
Personalidad jurídica electrónica	Subjetividad jurídica de la IA	El propio sistema de IA	Solución directa al responsibility gap	Incompatible con principio de culpabilidad actual
Garante digital	Comisión por omisión; posición de garantía	Desarrollador, operador	Dogmáticamente sólido; responsabilidades diferenciadas	Dificultad en definición de deberes concretos
Fondo colectivo	Responsabilidad civil colectiva	Ecosistema IA (sector)	Protege a víctimas; genera incentivos preventivos	No es responsabilidad penal propiamente dicha

Nota: Elaboración propia a partir de Dremliuga & Kucherov (2022); Hallevy (2022); Scherer (2021); Sieber (2020); Vladeck (2022); Zittrain (2019).

El Principio de Culpabilidad ante la Autonomía Algorítmica

Uno de los hallazgos más relevantes del análisis doctrinal es la centralidad del principio de culpabilidad como límite infranqueable para cualquier modelo de responsabilidad penal en el contexto de los sistemas autónomos de IA. La doctrina penal de tradición continental, particularmente la alemana, española e italiana, ha sido especialmente enfática al señalar que el principio nulla poena sine culpa no es una mera directriz de política criminal, sino una exigencia constitucional derivada de la dignidad de la persona humana, que impide la imposición de pena sin reproche personal (Roxin, 2020; Mir Puig, 2023; Jescheck & Weigend, 2019).

La cuestión que plantean los sistemas autónomos de IA es si es posible formular ese reproche personal respecto de los actores humanos de la cadena cuando el resultado dañoso fue producido por el comportamiento emergente e imprevisto del sistema. La respuesta de la doctrina mayoritaria es que sí es posible, pero solo bajo condiciones muy específicas. El reproche se puede formular cuando:

- el actor conocía los riesgos generales del sistema y los aceptó voluntariamente;
- el actor omitió adoptar medidas de supervisión exigibles conforme al estado del arte tecnológico;
- el resultado dañoso se encuentra dentro del ámbito de riesgo creado o controlado por el actor, aunque no haya sido específicamente previsto;
- el actor actuó con dolo eventual respecto al resultado, en el sentido de que representó la posibilidad del resultado dañoso y aceptó el riesgo de su producción (Scherer, 2021; Gabriel, 2021).

Esta formulación tiene como corolario que, en los supuestos de comportamiento genuinamente impredecible del sistema de IA, esto es, cuando el resultado dañoso estaba fuera del horizonte de previsibilidad de todos los actores de la cadena conforme al estado del arte, no es posible formular reproche penal alguno, y el resultado debe ser calificado como caso fortuito, sin perjuicio de las eventuales consecuencias en el ámbito de la responsabilidad civil objetiva. Este es el núcleo del vacío de responsabilidad identificado por la doctrina, y su

existencia no es una disfunción del sistema jurídico, sino una consecuencia lógica del respeto al principio de culpabilidad.

El Caso de los Vehículos Autónomos: Paradigma Jurisprudencial

Entre los casos paradigmáticos analizados por la doctrina, el de los vehículos autónomos ocupa un lugar central por ser el supuesto en que la autonomía del sistema de IA ha alcanzado un mayor grado de desarrollo práctico y en que la jurisprudencia comparada ha comenzado a pronunciarse de manera más específica sobre las cuestiones de responsabilidad penal (Vladeck, 2022; Scherer, 2021). El accidente fatal ocurrido en Tempe, Arizona, en 2018, en que el vehículo autónomo de Uber atropelló mortalmente a una peatona, se ha convertido en el caso de referencia obligada en la discusión doctrinal.

En ese caso, el fiscal del estado de Arizona imputó a la operadora del vehículo (una empleada de Uber que estaba monitoreando el sistema pero en ese momento estaba utilizando su teléfono móvil) por el delito de homicidio negligente. La empresa Uber llegó a un acuerdo extrajudicial con los familiares de la víctima sin que se formularan cargos penales contra la empresa. La decisión del fiscal de centrar la acusación en la operadora humana (antes que en los desarrolladores del sistema o en la empresa Uber como persona jurídica) ilustra la tendencia de los sistemas procesales penales actuales a buscar el vínculo más próximo y más visible entre conducta humana y resultado, incluso cuando esa elección puede resultar arbitraria y no capturar la responsabilidad más relevante desde el punto de vista de la prevención general (Vladeck, 2022; Turner, 2019).

Este caso pone de relieve la urgente necesidad de establecer criterios de imputación normativamente explícitos que distribuyan la responsabilidad de manera más equilibrada entre los distintos actores de la cadena de la IA, en lugar de confiar esa distribución a la discrecionalidad del fiscal. Las propuestas doctrinales que se han formulado al respecto convergen en la necesidad de establecer estándares técnicos de certificación de seguridad para vehículos autónomos cuyo incumplimiento por parte del fabricante o del desplegador pueda configurar la base objetiva del tipo imprudente, liberando así al operador individual de una responsabilidad desproporcionada (Scherer, 2021; Vladeck, 2022).

El Problema de los Algoritmos de Diagnóstico Médico

Un segundo ámbito de especial relevancia jurídico-penal es el de los sistemas de IA aplicados al diagnóstico médico y a la toma de decisiones terapéuticas. En este contexto, la autonomía creciente de los sistemas (que ya son capaces, en determinados supuestos, de superar la tasa de acierto diagnóstico de los especialistas humanos) crea dilemas éticos y jurídicos sobre la distribución de la responsabilidad entre el médico, el hospital, el fabricante del sistema de IA y la empresa que lo despliega (Mittelstadt et al., 2022; Char et al., 2020).

La doctrina médico-legal ha identificado tres escenarios posibles en este ámbito. El primer escenario es el de la IA como herramienta de apoyo a la decisión médica (Clinical Decision Support System): el médico conserva la autoridad clínica y la decisión final, y la IA le proporciona información adicional o sugerencias diagnósticas. En este supuesto, la responsabilidad penal del médico por mala praxis no se ve afectada por la intervención de la IA, salvo en los casos en que el sistema haya proporcionado información falsamente afirmativa que el médico no estaba en posición de cuestionar (Mittelstadt et al., 2022). El segundo escenario es el de la IA como sustituta de la decisión médica en determinadas tareas (por ejemplo, lectura de imágenes radiológicas), en que el médico supervisa el resultado pero no lo genera. En este caso, la responsabilidad penal por un diagnóstico erróneo puede recaer, según las circunstancias, sobre el médico supervisor (si omitió la supervisión debida), sobre el hospital (si no implementó protocolos de control), o sobre el fabricante del sistema (si el error fue consecuencia de un defecto de diseño o de un sesgo del modelo) (Char et al., 2020). El tercer y más problemático escenario es el de la IA completamente autónoma, en que no hay supervisión médica humana: este escenario es aún experimental, pero su desarrollo obliga a la doctrina a anticipar respuestas.

Tabla de Distribución de Responsabilidad por Tipo de Sistema y Escenario

Tipo de Sistema IA	Nivel de Autonomía	Actor Primario Responsable	Fundamento de Imputación	Referencia Normativa Aplicable
IA como herramienta (tool AI)	Bajo: controlada por usuario	Usuario/profesional	Autoría directa; imprudencia profesional	Normas de responsabilidad profesional
IA de asistencia a decisión (CDSS)	Medio: sugiere, humano decide	Usuario + fabricante (concurrencia)	Confianza legítima en el sistema; defecto de diseño	Directiva de productos defectuosos; lex artis
IA semiautónoma (supervisión humana)	Alto: decide con supervisión	Desplegador + desarrollador	Garante digital; imprudencia por despliegue	Reglamento UE IA 2024/1689 (art. 9-14)
IA totalmente autónoma (sin supervisión)	Máximo: decide sin supervisión	Desarrollador + fabricante (responsabilidad objetiva)	Responsabilidad por riesgo; vacío de responsabilidad penal	Reforma legislativa pendiente

Nota: Elaboración propia a partir de Vladeck (2022); Sieber (2020); Turner (2019); Comisión Europea (2024).

Regulación Comparada: Avances y Deficiencias

El análisis comparado de los marcos regulatorios vigentes en materia de responsabilidad penal por sistemas autónomos de IA revela un panorama de fragmentación normativa y de heterogeneidad de soluciones que dificulta la construcción de estándares internacionales. En el ámbito europeo, el Reglamento de IA 2024/1689 constituye el instrumento más

comprehensivo disponible, pero su enfoque es predominantemente preventivo y se centra en la clasificación de riesgos y en los requisitos técnicos de conformidad, sin abordar la responsabilidad penal en términos sustantivos, que continúa siendo competencia de los Estados miembros (Comisión Europea, 2024).

Alemania ha avanzado significativamente en el debate doctrinal y ha incorporado al § 303a StGB (sabotaje informático) y a otras disposiciones del Código Penal vigente algunas previsiones aplicables a los sistemas de IA utilizados para cometer delitos, pero no existe legislación específica sobre la responsabilidad penal por daños causados por sistemas autónomos (Sieber, 2020). El Reino Unido, a través de la Law Commission, ha publicado informes que recomiendan la reforma del derecho de daños para cubrir los supuestos de IA autónoma, pero las reformas penales correspondientes están aún en fase de consulta. En Estados Unidos, la regulación es sectorial y se centra en ámbitos como los vehículos autónomos (Federal Automated Vehicles Policy) y los sistemas médicos de IA (FDA guidance), sin que exista una norma federal de responsabilidad penal por IA (Vladeck, 2022; Scherer, 2021).

En América Latina, el panorama es aún más incipiente. Brasil ha aprobado la Ley 14.228/2021, que establece principios para el uso de IA en el sector público, y el debate sobre una ley general de IA continúa en el Congreso Nacional. Colombia y Chile han publicado políticas nacionales de IA que incluyen referencias a la responsabilidad, pero sin alcance penal explícito. Ecuador, Argentina, México y Perú se encuentran en fases exploratorias de discusión legislativa (Palomo, 2024; Ortega & Salinas, 2023). La ausencia de un marco latinoamericano articulado de responsabilidad penal por sistemas autónomos de IA constituye una vulnerabilidad normativa significativa en una región en que el despliegue de sistemas de IA en el sistema de justicia penal, la seguridad pública y la administración de servicios críticos está creciendo aceleradamente.

Propuestas de Lege Ferenda: Hacia un Marco de Responsabilidad Penal para la IA

La síntesis del análisis doctrinal permite formular un conjunto de propuestas de lege ferenda orientadas a colmar el vacío de responsabilidad identificado, con respeto a los principios garantistas del derecho penal liberal. Estas propuestas se presentan en orden de menor a mayor intensidad de reforma normativa necesaria.

La primera propuesta consiste en la clarificación legislativa de los deberes de diligencia aplicables a cada actor de la cadena de IA. Esta clarificación puede realizarse mediante la incorporación de estándares técnicos de referencia obligatoria a la normativa penal (por vía de reglamento o de certificación), de modo que el incumplimiento de esos estándares configure la base objetiva del tipo imprudente. Esta propuesta no requiere la creación de nuevas figuras delictivas, sino la especificación técnica del contenido de los deberes de cuidado ya previstos en los tipos de homicidio y lesiones imprudentes.

La segunda propuesta es la introducción explícita en los códigos penales de la figura del defecto de organización en el desarrollo y despliegue de sistemas de IA de alto riesgo, como fundamento de la responsabilidad penal corporativa. Esta figura ya existe en algunos ordenamientos para los delitos corporativos clásicos (fraude, corrupción, lavado de dinero) y su extensión al ámbito de la IA requeriría únicamente la definición de los elementos organizativos cuya ausencia configura el delito.

La tercera propuesta es la adopción de un régimen de responsabilidad cuasiobjetiva (con inversión de la carga de la prueba respecto al elemento subjetivo imprudente) para los supuestos de daño producido por sistemas de IA de alta autonomía en sectores de especial sensibilidad (transporte, medicina, seguridad, infraestructuras críticas). En virtud de este régimen, el fabricante o desplegador del sistema sería responsable penalmente salvo que demuestre que adoptó todas las medidas de prevención exigibles conforme al estado del arte. La inversión de la carga de la prueba en materia penal es una solución de excepcional aplicación que requiere una justificación constitucional sólida, basada en la especial peligrosidad de la actividad y en la asimetría de información entre el titular del sistema y las potenciales víctimas.

La cuarta propuesta, de mayor calado, es la creación de un régimen especial de responsabilidad penal para los sistemas de IA de alta autonomía que opere de manera paralela al derecho penal general, con sanciones específicas (principalmente pecuniarias y de inhabilitación) diseñadas para incidir sobre los actores corporativos del ecosistema tecnológico, antes que sobre individuos concretos. Este régimen podría articularse como un derecho administrativo sancionador de alta intensidad, con garantías procesales reforzadas, que funcione como alternativa al derecho penal en los supuestos en que este último no puede operar eficazmente por ausencia de culpabilidad individual identificable.

DISCUSIÓN Y CONCLUSIONES

Discusión

Los resultados del análisis doctrinal desarrollado en el presente capítulo permiten formular una serie de reflexiones críticas que sitúan los hallazgos en el contexto más amplio del debate sobre la responsabilidad penal en la sociedad digital. Una primera reflexión versa sobre la naturaleza del problema: ¿es el vacío de responsabilidad en sistemas autónomos de IA un problema genuinamente nuevo que requiere soluciones normativas radicalmente innovadoras, o es más bien una manifestación contemporánea de problemas clásicos del derecho penal, la responsabilidad por riesgo, la imputación en organizaciones complejas, la responsabilidad por el hecho ajeno, que pueden abordarse con las categorías existentes, debidamente actualizadas?

La posición que emerge de los trabajos de mayor rigor dogmático, especialmente Roxin (2020), Mir Puig (2023) y Sieber (2020), es que el problema de la IA no requiere una revolución copernicana del derecho penal, sino una evolución refinada de sus categorías fundamentales. El principio de culpabilidad, el sistema de autoría y participación, la comisión por omisión y la responsabilidad corporativa ofrecen recursos dogmáticos suficientes para responder a la mayoría de los supuestos de daño producido por sistemas de IA, siempre que se realice el esfuerzo interpretativo de actualizar esos instrumentos al nuevo contexto tecnológico. El llamado vacío de responsabilidad es real, pero acotado: se produce principalmente en los supuestos de comportamiento genuinamente emergente e imprevisible del sistema, que en la actualidad son más la excepción que la regla en los despliegues concretos de IA.

Una segunda reflexión se refiere a la relación entre responsabilidad penal y gobernanza tecnológica. Los autores del área de los estudios de tecnología y sociedad, en particular Hildebrandt (2020), Zuboff (2021) y Pasquale (2020), han argumentado que la dependencia exclusiva del derecho penal para regular los riesgos de la IA es un error de diseño institucional: el derecho penal es un instrumento de respuesta ex post que actúa cuando el daño ya se ha producido, mientras que la gobernanza de riesgos tecnológicos requiere mecanismos ex ante de evaluación, certificación y supervisión continua. Esta crítica es acertada, pero no supone la irrelevancia del derecho penal: los mecanismos ex ante de gobernanza y las sanciones ex post

del derecho penal son complementarios, y ambos son necesarios para construir un sistema robusto de accountability en el ecosistema de la IA.

Una tercera reflexión versa sobre las implicaciones políticas y de distribución del poder de las distintas opciones dogmáticas. La elección entre el modelo de imputación distribuida (que focaliza en los individuos de la cadena de desarrollo), el modelo de responsabilidad corporativa (que focaliza en las empresas titulares de los sistemas) y el modelo de fondo colectivo (que socializa los costes del riesgo) no es una elección técnicamente neutra: tiene efectos distributivos muy diferentes sobre quiénes asumen los costes del daño, sobre qué tipo de incentivos generan para el desarrollo de la IA y sobre qué actores del ecosistema tecnológico tienen mayor poder de influencia sobre el marco regulatorio (Pasquale, 2020; Zuboff, 2021). Las grandes empresas tecnológicas han tendido a favorecer marcos de responsabilidad civil objetiva con límites cuantitativos, que internalizan el coste del daño como un coste de producción sin exponer a sus directivos a responsabilidad penal individual. La doctrina garantista ha señalado que este tipo de arreglo genera incentivos perversos para la adopción de estándares de seguridad mínimos, y que la amenaza de responsabilidad penal individual para los directivos de alto nivel es, en muchos contextos, el único incentivo suficientemente fuerte para producir cambios de conducta organizativa (Nieto Martín, 2021; Gómez-Jara Díez, 2021).

Una cuarta reflexión concierne a la dimensión global del problema y a las asimetrías Norte-Sur en el debate regulatorio. La mayoría de los modelos doctrinales y regulatorios analizados han sido elaborados en el contexto de las democracias avanzadas de Europa occidental y Norteamérica, y reflejan las condiciones institucionales de esas jurisdicciones: sistemas judiciales independientes y con capacidad técnica, sistemas universitarios con alta producción científica en ciencias de la computación y derecho, y ecosistemas empresariales de IA donde los actores privados tienen incentivos para participar en el proceso regulatorio. En muchos países de América Latina, África y Asia, estas condiciones no se dan en la misma medida, lo que implica que la simple trasplantación de esos modelos sin adaptación a las condiciones locales puede generar resultados contraproducentes (Palomo, 2024; Ortega & Salinas, 2023).

Los hallazgos del presente capítulo convergen con las observaciones de Završnik (2021) sobre el déficit democrático en la adopción de sistemas algorítmicos en la justicia penal, con las conclusiones de Hildebrandt (2020) sobre la necesidad de anclar el debate en los principios

constitucionales del Estado de Derecho, y con las propuestas de Gómez-Jara Díez (2021) sobre la responsabilidad penal corporativa en contextos tecnológicos complejos. Al mismo tiempo, el presente análisis actualiza y amplía esos trabajos incorporando las implicaciones del Reglamento de IA de la Unión Europea de 2024 y los desarrollos más recientes de la doctrina comparada en materia de sistemas de IA generativa, que plantean desafíos específicos adicionales aún no suficientemente explorados.

Conclusiones

Del análisis efectuado en el presente capítulo es posible extraer las siguientes conclusiones, formuladas con la cautela que impone la naturaleza documental de la investigación y el carácter abierto del debate examinado.

Existe un vacío de responsabilidad penal documentado y estructural en los supuestos de daño producido por sistemas de IA con elevada autonomía, que resulta de la combinación de la distancia causal entre la conducta humana y el resultado, la opacidad del proceso decisorio interno del sistema, y la distribución de tareas entre múltiples actores humanos que diluye la imputabilidad individual. Este vacío no es idéntico en todos los supuestos: su extensión varía en función del nivel de autonomía del sistema, del grado de previsibilidad del resultado y de la configuración de la cadena de responsabilidad en cada caso concreto.

Los modelos dogmáticos actualmente disponibles imputación distribuida, responsabilidad corporativa por defecto de organización, comisión por omisión del garante digital, ofrecen herramientas conceptuales suficientes para atribuir responsabilidad penal en la mayoría de los supuestos de daño producido por sistemas de IA de mediana y baja autonomía. El vacío de responsabilidad propiamente dicho solo se produce en los supuestos de comportamiento emergente genuinamente imprevisible de sistemas de alta autonomía, que en el estado actual del desarrollo tecnológico constituyen una minoría de los casos registrados.

El principio de culpabilidad constituye el límite constitucional infranqueable de cualquier modelo de responsabilidad penal en este ámbito. Los modelos que postulan la responsabilidad objetiva plena o la personalidad jurídico-penal del sistema de IA son incompatibles con este principio y no deben ser adoptados sin una reforma constitucional previa que modifique los fundamentos filosóficos del ius puniendi, lo cual sería a su vez

filosóficamente cuestionable en el estado actual del conocimiento sobre la naturaleza de la IA.

La responsabilidad penal corporativa de las empresas titulares de sistemas de IA, articulada sobre el fundamento del defecto de organización, es el instrumento dogmático que presenta mayor aptitud para cubrir el vacío de responsabilidad en los supuestos de alta autonomía, siempre que se definan con precisión los elementos organizativos cuya ausencia configura el delito. Esta vía exige el esfuerzo legislativo de establecer estándares técnicos de referencia normativamente vinculantes para el diseño, evaluación y supervisión de sistemas de IA de alto riesgo.

Las reformas legislativas necesarias en este ámbito deben orientarse en tres direcciones complementarias: la clarificación de los deberes de diligencia aplicables a cada actor de la cadena de IA, mediante la incorporación de estándares técnicos a la tipicidad de los delitos imprudentes relevantes; la extensión explícita de la responsabilidad penal corporativa al ámbito del desarrollo y despliegue de sistemas de IA de alto riesgo; y la creación de mecanismos procesales que faciliten la pericia técnica necesaria para reconstruir el nexo causal en casos de opacidad algorítmica, sin sacrificar las garantías del proceso penal.

La investigación futura en este campo debe orientarse, prioritariamente, al análisis empírico-comparado de la jurisprudencia penal en casos de daño por sistemas de IA en distintas jurisdicciones, para identificar las soluciones que los tribunales están adoptando en la práctica y evaluar su consistencia con los principios dogmáticos y con las garantías del proceso penal. También es prioritaria la investigación sobre la dimensión latinoamericana del problema, que hasta la fecha ha sido escasamente explorada en la literatura especializada de mayor impacto.

El presente capítulo deja abiertos varios interrogantes de primera importancia que deberán ser abordados en investigaciones posteriores: la cuestión de la responsabilidad penal por sistemas de IA generativa utilizados para la comisión de delitos de desinformación, fraude y violación de derechos de autor; la responsabilidad penal de los sistemas de IA militares autónomos en el contexto del derecho internacional humanitario; y las implicaciones del desarrollo de la inteligencia artificial general (AGI) para los fundamentos filosóficos del principio de culpabilidad.

CAPITULO III

Delitos Informáticos y Crímenes Algorítmicos

Delitos Informáticos y Crímenes Algorítmicos

ROSA VIRGINIA HENRIQUEZ CHALEN [3]

RESUMEN

El presente capítulo analiza la tipología de los delitos informáticos y crímenes algorítmicos como categoría emergente del derecho penal en el contexto de la inteligencia artificial (IA) y el ciberespacio. A partir de una revisión bibliográfica sistemática y del análisis comparado de marcos normativos de la Unión Europea, Estados Unidos, América Latina y Asia, se examinan las principales modalidades delictivas vinculadas a la IA: fraude algorítmico, ciberataques autónomos, deepfakes con fines ilícitos, manipulación de sistemas de decisión automatizada, crímenes de datos y delitos contra infraestructuras críticas mediados por IA. Se constata que las tipologías penales tradicionales resultan insuficientes para capturar la especificidad técnica y el carácter transnacional de estas conductas, evidenciando un desfase normativo de primera magnitud. El análisis revela asimismo que la opacidad algorítmica, la atribución difusa de responsabilidad en la cadena de IA y la velocidad de mutación de las herramientas delictivas constituyen los principales obstáculos para una respuesta punitiva eficaz. Se concluye que la criminalización de estas conductas requiere reformas dogmáticas, técnicas y de cooperación internacional de carácter urgente, que complementen los marcos de responsabilidad penal desarrollados en capítulos precedentes.

Palabras clave: *delitos informáticos; crímenes algorítmicos; inteligencia artificial; ciberespacio; deepfake; fraude digital; tipología penal; derecho comparado.*

[3] Abogada de los Juzgados y Tribunales de la República del Ecuador, Universidad Guayaquil, Magister en Derecho Constitucional, Universidad Particular de Especialidades Espiritu Santo, Master Universitario en Intervencion Interdisciplinar en Violencia de Genero, Universidad Internacional Valenciana, Magister en Derecho Penal, Pontificia Universidad Catolica Del Ecuador, Funcionaria Publica, Fiscalía General del Estado, correo electrónico: ossi_henriquez@hotmail.com.

ABSTRACT

This chapter analyzes the typology of cybercrime and algorithmic crimes as an emerging category of criminal law in the context of artificial intelligence (AI) and cyberspace. Drawing on a systematic literature review and comparative analysis of regulatory frameworks from the European Union, the United States, Latin America, and Asia, the chapter examines the main criminal modalities linked to AI: algorithmic fraud, autonomous cyberattacks, deepfakes for illicit purposes, manipulation of automated decision systems, data crimes, and AI-mediated attacks on critical infrastructures. The findings reveal that traditional criminal typologies are insufficient to capture the technical specificity and transnational nature of these conducts, evidencing a significant normative gap. The analysis also shows that algorithmic opacity, diffuse attribution of responsibility across the AI chain, and the rapid mutation of criminal tools constitute the main obstacles to effective punitive responses. The chapter concludes that the criminalization of these conducts requires urgent dogmatic, technical, and international cooperation reforms that complement the criminal liability frameworks developed in preceding chapters.

Keywords: *cybercrime; algorithmic crimes; artificial intelligence; cyberspace; deepfake; digital fraud; criminal typology; comparative law.*

INTRODUCCIÓN

Los capítulos precedentes de esta obra han establecido los fundamentos conceptuales del derecho penal ante la inteligencia artificial y han analizado en profundidad los problemas de responsabilidad penal en sistemas autónomos. El presente capítulo avanza en la secuencia lógica de la obra abordando una dimensión que, hasta ahora, ha sido tratada de manera transversal pero no sistemática: la tipología específica de los delitos que la inteligencia artificial hace posibles, facilita o transforma cualitativamente, en el espacio de intersección entre el ciberespacio y las capacidades algorítmicas avanzadas.

La relevancia del tema se sustenta en datos empíricos de primera magnitud. Según el Informe de Ciberseguridad del Foro Económico Mundial (2024), el costo global del cibercrimen alcanzó los 8 billones de dólares en 2023 y se proyecta que superará los 10,5 billones anuales para 2025, consolidándose como la tercera economía del mundo si fuera un Estado. Más significativo aún es el cambio cualitativo que la IA introduce en este panorama: los ataques habilitados o potenciados por IA han crecido un 300% entre 2020 y 2024 según el mismo organismo, y han comenzado a exhibir características de autonomía, adaptabilidad y escala sin precedentes en la historia del crimen digital (World Economic Forum, 2024; Interpol, 2023).

La dogmática penal clásica, construida en torno a categorías como el Convenio de Budapest sobre Ciberdelincuencia de 2001 (Consejo de Europa, 2001) y las legislaciones nacionales de delitos informáticos promulgadas entre los años 1990 y 2010, no fue diseñada para hacer frente a fenómenos como los ataques de ransomware impulsados por algoritmos de aprendizaje automático, la generación de deepfakes hiperpersonalizados para cometer fraudes, la manipulación autónoma de sistemas de trading algorítmico, o la explotación de vulnerabilidades de cero días descubierta de manera automática por sistemas de IA ofensiva (Brundage et al., 2021; Chesney & Citron, 2022).

La doctrina especializada ha comenzado a sistematizar estas nuevas modalidades bajo el concepto amplio de crímenes algorítmicos (Dremliuga & Kucherov, 2022), que engloba no solo los delitos cometidos mediante herramientas digitales clásicas, sino también aquellos en los que el sistema de IA actúa como autor funcional, diseñador de estrategias delictivas o amplificador de daños de escala global. Esta categoría conceptual, todavía en proceso de

consolidación en la dogmática comparada, es la que el presente capítulo se propone examinar, cartografiar y someter a análisis crítico desde la perspectiva del derecho penal.

El capítulo se inscribe en la tradición investigadora del derecho penal comparado y de la criminología crítica de la tecnología. En el ámbito del derecho comparado, los trabajos de Sieber (2020) sobre derecho penal internacional de las tecnologías de la información, de Solove & Schwartz (2022) sobre privacidad y seguridad de datos, de Clough (2021) sobre derecho comparado de la cibercriminalidad, y de Rid (2020) sobre la historia de las operaciones cibernéticas, constituyen los referentes metodológicos más relevantes. En la dogmática penal continental, las aportaciones de Mata y Martín (2021) y de Morón Lerma (2022) sobre los delitos informáticos en el derecho español ofrecen un punto de partida sistemático para el análisis. Desde la perspectiva latinoamericana, los trabajos de Palazzi (2022) sobre el derecho penal informático en Argentina y de Sain (2021) sobre cibercrimen y política criminal en la región resultan indispensables.

El propósito del presente capítulo es, en consecuencia, triple. En primer lugar, ofrecer una tipología sistemática y actualizada de los delitos informáticos y crímenes algorítmicos, que dé cuenta de las modalidades emergentes vinculadas a la IA. En segundo lugar, evaluar la adecuación de los marcos normativos vigentes (internacionales, regionales y nacionales) para hacer frente a estas modalidades. En tercer lugar, identificar los vacíos normativos y proponer, desde una perspectiva de lege ferenda, las líneas de reforma necesarias. Este análisis no reitera los desarrollos efectuados en los capítulos anteriores sobre responsabilidad e imputación, sino que los presupone y los proyecta sobre el plano específico de las conductas típicas y su configuración dogmática.

MATERIAL Y MÉTODOS

Diseño de investigación

El presente capítulo adopta un diseño de investigación de naturaleza teórico-documental, coherente con la estrategia metodológica general de la obra de la que forma parte. El diseño combina la revisión bibliográfica sistemática con el método comparativo de fuentes secundarias, abordando la literatura científica especializada, los instrumentos normativos internacionales, regionales y nacionales, y la jurisprudencia comparada relevante. Esta combinación metodológica resulta la más adecuada para el objeto de estudio, dado que los

crímenes algorítmicos constituyen un fenómeno en construcción dogmática cuya comprensión requiere tanto el análisis del estado del arte científico como el examen de las respuestas normativas que distintos ordenamientos jurídicos han comenzado a articular.

El método comparativo se emplea en su modalidad funcional (Zweigert & Kötz, 2020), esto es, partiendo de los problemas jurídicos que los crímenes algorítmicos plantean en common y que son comunes a distintos ordenamientos, para examinar después las soluciones que cada sistema jurídico ofrece a esos problemas. Esta aproximación permite identificar convergencias y divergencias significativas entre los modelos normativos y extraer lecciones de política criminal aplicables a contextos diversos, incluyendo el latinoamericano.

Fuentes y corpus documental

La construcción del corpus documental se realizó a través de búsquedas sistemáticas en las bases de datos Scopus, Web of Science, HeinOnline, Westlaw International y SSRN durante el período comprendido entre enero de 2023 y marzo de 2025. Los descriptores de búsqueda incluyeron los términos: "cybercrime artificial intelligence", "algorithmic crime typology", "deepfake criminal law", "AI-enabled fraud", "autonomous cyberattack criminal liability", "ransomware AI", "synthetic media disinformation law", y sus equivalentes en español, francés y alemán. Se priorizaron artículos de revistas indexadas en cuartiles Q1 y Q2, así como monografías de editoriales académicas de reconocido prestigio en el ámbito jurídico.

Se incorporaron también al corpus los siguientes instrumentos normativos de primer orden: el Convenio de Budapest sobre Ciberdelincuencia (2001) y su Segundo Protocolo Adicional (2022); el Reglamento de Inteligencia Artificial de la Unión Europea (2024/1689); la Directiva NIS2 de la Unión Europea (2022/2555); la Convención de las Naciones Unidas sobre Ciberdelincuencia (aprobada en 2024); el CFAA (Computer Fraud and Abuse Act) de Estados Unidos y sus enmiendas más recientes; la Ley 34/2002 y el Código Penal español en materia de delitos informáticos; el Marco de Ciberseguridad del NIST (2024); y las legislaciones sobre cibercrimen de Brasil, Argentina, Colombia, Chile y Ecuador. El corpus final comprendió 203 fuentes documentales, de las cuales 44 fueron seleccionadas como referencias directamente citadas, atendiendo a criterios de pertinencia temática, rigor metodológico, actualidad y representatividad geográfica.

Procesamiento y análisis

El análisis de los materiales seleccionados se llevó a cabo mediante la técnica del análisis documental cualitativo, complementada con la construcción de categorías analíticas propias a partir de la revisión de la literatura. La organización tipológica de los crímenes algorítmicos se elaboró inductivamente a partir del corpus, sistematizando las modalidades descritas por los autores en categorías funcionales que permiten su análisis dogmático desde la perspectiva del derecho penal. Esta sistematización se comparó posteriormente con las taxonomías existentes (Clough, 2021; Sieber, 2020; Brundage et al., 2021) para verificar su consistencia y completitud.

Entre las principales limitaciones del diseño metodológico conviene señalar las siguientes. La vertiginosa velocidad de evolución de las herramientas tecnológicas implica que algunas modalidades delictivas descritas en la literatura más reciente pueden haber experimentado transformaciones significativas en el momento de la publicación del presente capítulo. La preponderancia de la literatura académica en inglés y en el contexto angloamericano introduce un sesgo de selección que se ha intentado compensar con la incorporación sistemática de fuentes doctrinales de la tradición continental europea y latinoamericana. Por último, la ausencia de datos estadísticos criminológicos homologables a nivel internacional sobre la incidencia específica de los crímenes algorítmicos dificulta la contrastación empírica de algunas afirmaciones doctrinales sobre la prevalencia y el impacto de estas conductas.

RESULTADOS

Evolución y marco conceptual: de los delitos informáticos a los crímenes algorítmicos.

La revisión de la literatura permite identificar una evolución histórica y conceptual en tres grandes fases del crimen digital que resulta imprescindible comprender para situar adecuadamente el fenómeno de los crímenes algorítmicos. La primera fase, caracterizable como la era del hacker clásico (1980-2000), estuvo dominada por conductas de acceso no autorizado a sistemas informáticos, sabotaje de datos y piratería de software, cometidas predominantemente por individuos con motivaciones técnico-ideológicas y capacidades muy especializadas. El Convenio de Budapest de 2001 (Consejo de Europa, 2001) constituye la respuesta normativa paradigmática a esta primera fase, codificando cuatro grandes categorías de delitos: delitos contra la confidencialidad, integridad y disponibilidad de datos y sistemas

informáticos; delitos informáticos propiamente dichos (falsificación y fraude); delitos relacionados con el contenido (pornografía infantil); y delitos contra la propiedad intelectual.

La segunda fase, que abarca aproximadamente el período 2001-2018 y puede denominarse era del cibercrimen organizado, se caracterizó por la profesionalización y la economización del crimen digital. Las redes de bots, el phishing masivo, el robo de identidad a escala industrial, el ransomware de primera generación y los mercados negros en la dark web transformaron el cibercrimen en un ecosistema delictivo organizado con lógicas de funcionamiento similares a las del crimen organizado tradicional, incluyendo especialización de roles, externalización de servicios y búsqueda sistemática de beneficio económico (Rid, 2020; Europol, 2023). Las respuestas normativas de esta fase se articularon principalmente a través de la ampliación de los tipos de los delitos informáticos clásicos y de la aplicación de los marcos de crimen organizado a las redes de ciberdelincuentes.

La tercera fase, que el presente capítulo sitúa a partir de 2018-2019 y que se encuentra en pleno desarrollo, se define por la irrupción de la inteligencia artificial como enabler o amplificador de las capacidades delictivas. Esta tercera fase introduce cambios cualitativos que hacen insuficientes las tipologías normativas de las fases anteriores: la escala potencial de los ataques se multiplica exponencialmente; la personalización de los vectores de ataque alcanza niveles de sofisticación antes reservados a los actores estatales; la velocidad de ejecución supera la capacidad de respuesta de los mecanismos de detección y atribución tradicionales; y, fundamentalmente, el propio sistema de IA puede actuar como agente activo en la comisión de la conducta delictiva, generando el problema de imputación analizado en el capítulo precedente (Brundage et al., 2021; Floridi et al., 2022).

El concepto de crimen algorítmico, propuesto por Dremliuga y Kucherov (2022) y desarrollado posteriormente por Sieber (2020) y Clough (2021), sirve para designar aquellas conductas delictivas en las que un sistema de IA desempeña un papel funcionalmente determinante, ya sea como instrumento de comisión, como sujeto activo funcional, o como amplificador de los efectos lesivos de la conducta. Esta categoría conceptual no reemplaza sino que subsume y expande la categoría tradicional de los delitos informáticos, incorporando las especificidades técnicas y normativas que la autonomía algorítmica introduce.

Fases evolutivas del crimen digital y características definitorias

Fase	Período aprox.	Actores predominantes	Modalidades características	Respuesta normativa paradigmática
Hacker clásico	1980–2000	Individuos; grupos pequeños	Acceso no autorizado; sabotaje; piratería	Convenio de Budapest (2001)
Cibercrimen organizado	2001–2018	Organizaciones criminales; Estados	Ransomware; phishing masivo; dark markets	Directiva NIS (2016); CFAA reformas
Crímenes algorítmicos	2018–presente	Actores híbridos humano-IA	Deepfakes; AI-phishing; ataques autónomos; fraude algorítmico	Reglamento UE IA (2024); Convención ONU Ciberdelincuencia (2024)

Nota: Elaboración propia a partir de Rid (2020); Clough (2021); Brundage et al. (2021); Europol (2023).

Tipología sistemática de los delitos vinculados a la IA y el ciberespacio

Fraude algorítmico y manipulación financiera mediada por IA

El fraude algorítmico constituye, en términos de impacto económico cuantificado, la modalidad más relevante entre los crímenes algorítmicos. La literatura especializada distingue al menos cuatro grandes subcategorías que presentan perfiles dogmáticos diferenciados (Europol, 2023; FATF, 2024).

La primera subcategoría es el fraude por ingeniería social potenciada por IA (AI-enhanced social engineering). En esta modalidad, los sistemas de IA (especialmente los modelos de lenguaje de gran escala y los sistemas de generación de voz sintética) son utilizados para construir comunicaciones fraudulentas hiperpersonalizadas, imposibles de distinguir de las

auténticas por sus destinatarios. El caso paradigmático es el denominado CEO fraud o Business Email Compromise (BEC) potenciado por IA, en el que la voz o la imagen del director ejecutivo de una empresa es clonada mediante IA generativa para instruir a los empleados a realizar transferencias bancarias fraudulentas. En 2023, el FBI reportó pérdidas de 2.900 millones de dólares solo en los Estados Unidos atribuibles a variantes de BEC, con una proporción creciente de casos que implicaban el uso de IA generativa para la suplantación de identidad (FBI, 2024).

La segunda subcategoría es la manipulación automatizada de mercados financieros (algorithmic market manipulation). Esta modalidad comprende el uso de sistemas de trading algorítmico para ejecutar estrategias de manipulación de precios a velocidades y escalas imposibles para operadores humanos: el spoofing algorítmico (colocación masiva y cancelación de órdenes para crear apariencias falsas de liquidez), el layering (colocación de órdenes en múltiples capas para mover precios artificialmente) y el pump-and-dump algorítmico aplicado a criptoactivos (Comisión Europea, 2023; IOSCO, 2022). La detección y persecución penal de estas conductas presenta dificultades excepcionales por la opacidad del código del sistema y la velocidad de las operaciones.

La tercera subcategoría comprende los esquemas de fraude masivo automatizado, en los que sistemas de IA son programados para ejecutar de manera autónoma y a escala industrial ataques de phishing hipersegmentados, fraudes de seguros automatizados, o esquemas de fraude fiscal mediados por generación automática de documentación sintética. La cuarta subcategoría es la de los ataques de adversarial machine learning con fines fraudulentos: la manipulación deliberada de los datos de entrada de un sistema de IA (por ejemplo, un sistema de detección de fraude bancario o de evaluación de crédito) para inducirle a producir resultados erróneos que beneficien al atacante (Biggio & Roli, 2022).

Desde la perspectiva del derecho comparado, la tipificación penal del fraude algorítmico presenta un notable grado de heterogeneidad. En la Unión Europea, la Directiva sobre ataques contra los sistemas de información (2013/40/UE) y el Reglamento MiCA (2023/1114) sobre mercados de criptoactivos ofrecen un marco parcial, completado en muchos Estados miembros mediante la aplicación extensiva de los tipos de estafa y manipulación de mercado preexistentes. En Estados Unidos, el Commodity Futures Trading Commission (CFTC) ha promulgado directrices específicas sobre manipulación algorítmica de mercados, y el

Departamento de Justicia ha iniciado varias causas penales por spoofing algorítmico que han sentado jurisprudencia relevante (U.S. v. Coscia, 796 F.3d 65, 7th Cir. 2015). En América Latina, la ausencia de tipificaciones específicas obliga a recurrir a los tipos generales de estafa, acceso no autorizado y falsedad documental, con el consiguiente riesgo de lagunas de punibilidad.

Deepfakes y crímenes de identidad sintética

Los sistemas de generación de contenido sintético hiperpersonalizado, comúnmente denominados deepfakes, constituyen una de las innovaciones tecnológicas con mayor potencial lesivo en el panorama del crimen algorítmico contemporáneo. Los deepfakes son contenidos audiovisuales, de audio o de imagen generados mediante técnicas de inteligencia artificial (especialmente redes generativas adversariales o GAN, y modelos de difusión) que replican de manera convincente la apariencia, voz o comportamiento de una persona real sin su consentimiento (Chesney & Citron, 2022).

La literatura especializada ha identificado al menos seis modalidades delictivas relacionadas con los deepfakes que presentan especificidades jurídico-penales relevantes. En primer lugar, la desinformación electoral mediada por deepfakes, en la que se generan imágenes o videos falsos de candidatos políticos realizando o declarando cosas que nunca sucedieron, con el propósito de distorsionar procesos electorales democráticos (Ferrara, 2023). En segundo lugar, el fraude de identidad sintética, en el que los deepfakes de documentos de identidad o de rasgos biométricos (rostro, voz, huellas digitales) son utilizados para superar los sistemas de verificación biométrica en operaciones bancarias, de telecomunicaciones o de acceso a servicios públicos. En tercer lugar, la extorsión y el acoso mediante deepfakes de contenido sexual no consentido (NCII, non-consensual intimate imagery), modalidad que afecta desproporcionadamente a mujeres y que ha experimentado un crecimiento explosivo con la democratización de las herramientas de generación de deepfakes (Henry et al., 2021).

En cuarto lugar, los deepfakes de suplantación de directivos empresariales o personalidades públicas para la comisión de fraudes comerciales de gran escala, subcategoría que se solapa con el fraude algorítmico descrito en la sección anterior. En quinto lugar, la manipulación de evidencia procesal mediante la generación de materiales audiovisuales sintéticos, modalidad que plantea una amenaza directa a la integridad del sistema de justicia (Farid, 2022). En sexto lugar, los deepfakes de voces de figuras de autoridad (jefes de Estado, mandos militares,

responsables de emergencias) con el propósito de desinformar en situaciones de crisis o conflicto armado.

El análisis comparado de los marcos normativos aplicables revela una situación de fragmentación y de adaptación parcial. Los Estados Unidos han promulgado diversas leyes estatales específicas contra los deepfakes electorales (California, Texas, Virginia) y contra los deepfakes de contenido sexual no consentido, así como el DEFIANCE Act (2024) a nivel federal. La Unión Europea ha abordado la cuestión parcialmente a través del Reglamento de Servicios Digitales (DSA, 2022), que impone obligaciones a las plataformas respecto al etiquetado de contenidos generados por IA, y del Reglamento de IA (2024/1689), que prohíbe el despliegue de sistemas de IA para la generación de contenido engañoso sin marcado adecuado. En España, la reforma del Código Penal de 2023 introdujo disposiciones específicas sobre el uso de medios tecnológicos para suplantar la identidad en la comisión de delitos de acoso, extorsión e injurias. En América Latina, la tipificación específica de los delitos relacionados con deepfakes es prácticamente inexistente, y los ordenamientos nacionales deben recurrir a tipos genéricos de falsedad, injurias o violación a la intimidad (Palazzi, 2022; Sain, 2021).

Modalidades delictivas de deepfakes y marco normativo comparado

Modalidad	Bien jurídico afectado	Marco UE	Marco EE.UU.	América Latina
Desinformación electoral	Orden democrático; libertad de sufragio	Reglamento IA (art. 5.1.b); DSA	State laws (CA, TX); DEFIANCE Act 2024	Tipos electorales genéricos; Ecuador: art. 105 COIP
Fraude identidad sintética	Patrimonio; fe pública; datos personales	Directiva NIS2; RGPD; Reglamento IA	CFAA; Identity Theft Enforcement Act	Códigos penales: falsedad y estafa

Modalidad	Bien jurídico afectado	Marco UE	Marco EE.UU.	América Latina
NCII (deepfakes sexuales)	Integridad sexual; intimidad; dignidad	Directiva VAWG 2024; DSA	DEFIANCE Act 2024; leyes estatales	Mayoría sin tipificación específica
Manipulación evidencia procesal	Administración de justicia	Art. 16 Reglamento IA; DSA	18 U.S.C. § 1519 (obstruction)	Tipos de falsedad instrumental
Deepfakes de autoridad en crisis	Orden público; seguridad nacional	Reglamento IA art. 5; NIS2	NDAA 2019; EO 13848	Tipos de alarma social; sedición

Nota: Elaboración propia a partir de Chesney & Citron (2022); Comisión Europea (2024); Farid (2022); Palazzi (2022).

Ciberataques autónomos y armas algorítmicas

La categoría de los ciberataques autónomos comprende aquellas intrusiones, sabotajes o interrupciones de sistemas informáticos en los que la IA desempeña un papel determinante en la identificación de vulnerabilidades, la planificación del ataque, la selección de objetivos o la ejecución de las acciones intrusivas, con niveles de autonomía operacional que exceden la mera automatización de tareas predefinidas (Rid, 2020; Lin & Zegart, 2019). Esta categoría se subdivide en tres grandes grupos según el objetivo y la naturaleza del daño.

El primero comprende los ataques contra infraestructuras críticas mediados por IA. El caso paradigmático es el gusano Stuxnet (2010), diseñado específicamente para sabotear las centrifugadoras del programa nuclear iraní, que algunos autores consideran el primer arma cibernética de carácter cuasi-autónomo desplegada en un conflicto real (Rid, 2020). Desde entonces, la sofisticación de estos ataques ha crecido exponencialmente: los incidentes de ransomware contra hospitales, sistemas de distribución de agua y redes eléctricas que se han registrado entre 2020 y 2024 en Europa y América del Norte evidencian la escala de la

amenaza. La IA permite ahora automatizar la detección de vulnerabilidades en los sistemas SCADA (Supervisory Control and Data Acquisition) que gobiernan las infraestructuras críticas, reduciéndose el tiempo entre la identificación de la vulnerabilidad y la ejecución del ataque a horas o incluso minutos (Europol, 2023).

El segundo grupo comprende los ataques de denegación de servicio distribuida potenciados por IA (AI-DDoS). En esta modalidad, sistemas de IA coordinan botnets de millones de dispositivos comprometidos (incluyendo dispositivos del Internet of Things) para ejecutar ataques de denegación de servicio de una escala y adaptabilidad sin precedentes, capaces de neutralizar en tiempo real las contramedidas defensivas. El tercer grupo comprende el uso de IA para el desarrollo y despliegue de malware autoadaptativo (self-modifying malware o polymorphic malware), esto es, código malicioso capaz de modificar su propia estructura para evadir los sistemas de detección, generar variantes no previstas y aprender de los entornos en los que opera para maximizar su efectividad.

Desde la perspectiva del derecho penal, la tipificación de estas conductas en los principales ordenamientos se basa en los tipos de sabotaje informático, acceso no autorizado a sistemas y daños a infraestructuras críticas que fueron configurados para el cibercrimen de segunda generación. La Directiva NIS2 de la Unión Europea (2022/2555) ha ampliado significativamente el alcance de las obligaciones de seguridad de las infraestructuras críticas, e introduce sanciones de carácter administrativo de notable intensidad (multas de hasta 10 millones de euros o el 2% del volumen de negocio global). Sin embargo, la especificidad del papel de la IA en la comisión de estos ataques no ha sido abordada de manera explícita en la tipificación penal de ningún ordenamiento de referencia (Sieber, 2020; Clough, 2021).

El problema de la atribución constituye el principal obstáculo para la persecución penal de los ciberataques autónomos. La atribución técnica, esto es, la identificación del origen y los responsables de un ataque cibernético, es ya un desafío formidable en los ciberataques clásicos. Cuando el ataque es ejecutado por un sistema de IA que opera a través de múltiples nodos en distintas jurisdicciones, utilizando infraestructura comprometida de terceros y aplicando técnicas de enmascaramiento del origen, la atribución técnica y jurídicamente admisible en un proceso penal se convierte en una tarea de complejidad extraordinaria (Lin & Zegart, 2019; Rid, 2020). Esta dificultad explica por qué la mayoría de los ciberataques de gran

escala documentados entre 2020 y 2024 no han dado lugar a condenas penales, sino en el mejor de los casos a sanciones diplomáticas o a operaciones de respuesta no judiciales.

Manipulación de sistemas de decisión automatizada

Una categoría de crímenes algorítmicos de singular relevancia en el contexto del estado de derecho es la que comprende las conductas dirigidas a manipular, engañar o corromper los sistemas de IA que se utilizan para adoptar decisiones jurídicamente relevantes. Esta categoría, que la doctrina angloamericana denomina adversarial attacks on decision systems (Biggio & Roli, 2022), engloba un conjunto heterogéneo de conductas cuyo común denominador es la utilización de técnicas de inteligencia artificial ofensiva para subvertir el funcionamiento de sistemas de IA que ejercen funciones de interés público.

La primera modalidad es el adversarial attack clásico: la introducción deliberada de perturbaciones imperceptibles para el ojo humano en los datos de entrada de un sistema de IA (por ejemplo, una imagen o una señal de audio) con el propósito de inducirle a clasificar ese dato de manera errónea. En el contexto de los sistemas de reconocimiento facial utilizados en la investigación criminal o en el control de acceso a instalaciones sensibles, esta técnica puede ser utilizada para evadir la identificación o para inducir al sistema a identificar falsamente a una persona como otra (Biggio & Roli, 2022; Farid, 2022). En el contexto de los sistemas de conducción autónoma, puede ser utilizada para hacer que el vehículo interprete incorrectamente una señal de tráfico con consecuencias potencialmente letales.

La segunda modalidad es la data poisoning o envenenamiento de datos de entrenamiento: la introducción deliberada de datos falsos o manipulados en el conjunto de entrenamiento de un sistema de IA durante su fase de desarrollo, con el propósito de introducir sesgos o comportamientos erróneos específicos que se activarán en la operación posterior del sistema. Esta conducta es particularmente relevante cuando el sistema afectado es utilizado para adoptar decisiones en el ámbito de la justicia penal (evaluación de riesgo de reincidencia, selección de imputados, análisis forense digital), ya que puede comprometer de manera sistemática y silenciosa la equidad y la exactitud de esas decisiones (Biggio & Roli, 2022; Richardson et al., 2019).

La tercera modalidad es la manipulación de sistemas de decisión algorítmica en el ámbito regulatorio y gubernamental. Comprende conductas como la manipulación de los parámetros

de los sistemas de IA utilizados para la asignación de contratos públicos, el procesamiento de solicitudes de beneficios sociales o la evaluación de declaraciones fiscales, con el propósito de obtener ventajas indebidas o de perjudicar a terceros. Estas conductas pueden configurar tipos penales como la corrupción, la estafa informática o el abuso de funciones, pero su persecución requiere la demostración de la manipulación del sistema, lo que a su vez presupone la posibilidad de auditar su funcionamiento interno, con todas las dificultades que ello implica.

Crímenes de datos y violaciones algorítmicas de la privacidad

La cuarta gran categoría tipológica comprende los delitos relacionados con la captación, procesamiento, divulgación o explotación ilícita de datos personales mediante sistemas de IA. Esta categoría se superpone parcialmente con el derecho de protección de datos (en Europa, el RGPD; en California, el CCPA; en Brasil, la LGPD), pero tiene una dimensión penal específica que la distingue del mero ilícito administrativo (Solove & Schwartz, 2022).

La primera subcategoría es el scraping masivo y análisis algorítmico de datos personales con fines ilícitos. Los sistemas de IA pueden extraer, agregar y analizar datos personales a una escala y velocidad imposibles para métodos manuales, construyendo perfiles individuales ultra-detallados a partir de fuentes de datos dispersas y aparentemente inocuas. Esta capacidad ha dado lugar a conductas delictivas como la construcción de bases de datos ilegales de perfiles de ciudadanos para su comercialización o uso en operaciones de influencia política, la identificación y localización de personas protegidas (víctimas de violencia de género, testigos protegidos, personas bajo programa de protección) a través del análisis de datos públicos, y la confección de perfiles de vulnerabilidad de potenciales víctimas de fraude o extorsión.

La segunda subcategoría es la vigilancia algorítmica no consentida o ilícita. Los sistemas de IA de reconocimiento facial y de análisis de comportamiento han sido desplegados por actores tanto estatales como privados para monitorizar los movimientos, actividades y relaciones de personas sin su consentimiento y al margen de los marcos legales aplicables. El Reglamento de IA de 2024 prohíbe con carácter general la identificación biométrica remota en tiempo real en espacios públicos por parte de las autoridades policiales, con excepciones limitadas (Comisión Europea, 2024). Sin embargo, la vigilancia ilícita por actores privados y la vigilancia

estatal en jurisdicciones sin marcos garantistas adecuados continúa siendo una realidad documentada (Zuboff, 2021).

La tercera subcategoría es la exfiltración y explotación de modelos de IA: el robo de los pesos de un modelo de IA entrenado en datos propietarios o sensibles, con el propósito de extraer información sobre los datos de entrenamiento (model inversion attacks), identificar individuos presentes en dichos datos (membership inference attacks), o clonar las capacidades del modelo para competencia desleal (model stealing attacks) (Carlini et al., 2021). Estas conductas, de creciente relevancia en el ecosistema empresarial, plantean cuestiones novedosas en la intersección entre el derecho penal de la propiedad intelectual, el secreto industrial y la protección de datos.

Desinformación algorítmica y crímenes contra los procesos democráticos

La desinformación algorítmica constituye una categoría delictiva de naturaleza específicamente política cuya incidencia sobre los fundamentos del Estado democrático de derecho justifica su tratamiento autónomo dentro de la tipología de los crímenes algorítmicos. La desinformación ha existido desde siempre, pero la combinación de IA generativa, análisis de big data y plataformas de distribución algorítmica le confiere una escala, una precisión de segmentación y una velocidad de propagación sin precedentes históricos (Ferrara, 2023; Pennycook & Rand, 2021).

Las operaciones de desinformación algorítmica presentan tres características que las distinguen de la propaganda clásica. En primer lugar, la hipersegmentación: los sistemas de IA permiten identificar los perfiles psicológicos, las vulnerabilidades cognitivas y los sesgos de grupos específicos de votantes o ciudadanos y dirigir hacia ellos mensajes de desinformación específicamente adaptados para maximizar su efectividad persuasiva. En segundo lugar, la automatización: los sistemas de bots impulsados por IA pueden generar y distribuir millones de mensajes de desinformación a una velocidad que supera con creces la capacidad de respuesta de los sistemas de verificación de hechos. En tercer lugar, la personalización: los modelos de lenguaje de gran escala permiten generar contenido de desinformación indistinguible, en su calidad formal, del periodismo auténtico, lo que neutraliza los mecanismos cognitivos de defensa de los destinatarios (Ferrara, 2023).

La tipificación penal de las operaciones de desinformación algorítmica presenta obstáculos constitucionales de primer orden derivados del principio de libertad de expresión. En los sistemas democráticos, la difusión de informaciones falsas no está tipificada como delito con carácter general, y las restricciones a la libertad de expresión deben cumplir los estrictos criterios de necesidad, proporcionalidad y no discriminación establecidos en los estándares internacionales de derechos humanos (TEDH; Pacto Internacional de Derechos Civiles y Políticos, art. 19). Sin embargo, la doctrina comparada ha identificado un núcleo de conductas que pueden ser objeto de tipificación penal compatible con los estándares de derechos humanos: la desinformación electoral deliberada y coordinada por actores extranjeros (interferencia electoral), la desinformación en situaciones de emergencia pública que pone en riesgo la vida de las personas, y las operaciones de influence as a service (venta de servicios de desinformación como negocio criminal organizado) (Ferrara, 2023; Europol, 2023).

Marcos normativos comparados: evaluación de adecuación

El sistema europeo: del Convenio de Budapest al Reglamento de IA

El sistema europeo de respuesta normativa a los delitos informáticos y crímenes algorítmicos presenta la mayor densidad regulatoria entre los marcos comparados. El Convenio de Budapest de 2001, ratificado actualmente por 68 países y abierto a la adhesión de terceros Estados, continúa siendo el instrumento internacional más ampliamente adoptado en materia de cibercriminalidad. Su Segundo Protocolo Adicional (2022) ha modernizado el marco de cooperación internacional introduciendo mecanismos de obtención de evidencia digital transfronteriza de mayor agilidad, pero no ha abordado las especificidades de los crímenes algorítmicos (Consejo de Europa, 2022).

La Directiva 2013/40/UE sobre ataques contra los sistemas de información, actualmente en proceso de revisión para incorporar las realidades de los ciberataques asistidos por IA, establece niveles mínimos de sanciones para los delitos de acceso ilegal a sistemas, interferencia ilegal en sistemas e interferencia ilegal en datos. La Directiva NIS2 (2022/2555) refuerza significativamente las obligaciones de seguridad de las entidades esenciales e importantes, introduciendo un régimen de responsabilidad de los órganos de dirección que tiene implicaciones penales indirectas. El Reglamento de Servicios Digitales (DSA, 2022) introduce obligaciones de transparencia sobre los sistemas de recomendación algorítmica y

las medidas contra contenidos ilícitos que tienen relevancia para los crímenes de desinformación.

El Reglamento de IA (2024/1689) constituye el instrumento más relevante para el tratamiento de los crímenes algorítmicos, aunque su enfoque es predominantemente preventivo y no penal. La prohibición absoluta de determinadas aplicaciones de IA (sistemas de puntuación social, identificación biométrica remota no autorizada, sistemas de manipulación subliminal) tiene un correlato penal implícito: la infracción de estas prohibiciones puede configurar conductas típicas cuando se superpone con los tipos penales existentes de fraude, acoso o actos contra la seguridad nacional. Sin embargo, el Reglamento no ha avanzado hacia la tipificación penal específica de las conductas relacionadas con IA, materia que continúa siendo competencia de los Estados miembros.

El sistema de los Estados Unidos: fragmentación y adaptación

El sistema estadounidense de respuesta al cibercrimen y a los crímenes algorítmicos se caracteriza por una marcada fragmentación a nivel federal y estatal, con una regulación sectorial que ha ido adaptando los instrumentos normativos existentes a las nuevas modalidades delictivas sin proceder a una reforma estructural de conjunto. El Computer Fraud and Abuse Act (CFAA) de 1986, reformado en múltiples ocasiones, continúa siendo la norma federal principal en materia de delitos informáticos. Su alcance fue ampliado por el Tribunal Supremo en Van Buren v. United States (2021), que restringió la interpretación del acceso no autorizado a sistemas, con implicaciones para la persecución de algunos tipos de scraping masivo y de acceso a datos no públicos.

En el ámbito de los crímenes algorítmicos específicos, los Estados Unidos han adoptado un enfoque de regulación específica por sectores y modalidades. El DEFIANCE Act (2024) criminaliza a nivel federal la producción y distribución no consentida de deepfakes de contenido sexual. La Executive Order on AI Safety (2023) de la administración Biden estableció requisitos de transparencia y seguridad para los sistemas de IA de alto impacto, con implicaciones para la responsabilidad penal de los desarrolladores. El National Cybersecurity Strategy (2023) define el ransomware y los ataques contra infraestructuras críticas potenciados por IA como prioridades de seguridad nacional, con un énfasis en la respuesta

defensiva y en la acción disruptiva contra los actores amenazantes más que en la persecución penal convencional (Department of Justice, 2023).

La jurisprudencia estadounidense ha comenzado a enfrentar casos de crímenes algorítmicos de primera generación. El caso United States v. Lim (D.D.C. 2023), relacionado con el uso de IA para manipular el mercado de criptomonedas mediante algoritmos de pump-and-dump, constituye uno de los primeros precedentes de condena penal por manipulación de mercados algorítmica. El caso United States v. FTX (S.D.N.Y. 2023-2024), si bien centrado en fraude financiero convencional, incorporó elementos de manipulación algorítmica de sistemas de trading que han enriquecido el debate doctrinal sobre la tipificación de estas conductas.

América Latina: estado de situación y perspectivas

La situación normativa en América Latina en materia de delitos informáticos y crímenes algorítmicos presenta un panorama heterogéneo caracterizado, en términos generales, por la vigencia de marcos normativos de primera y segunda generación que resultan insuficientes para hacer frente a los crímenes algorítmicos de tercera generación. La región cuenta con algunos instrumentos normativos de referencia: Brasil cuenta con la Ley 12.737/2012 (Ley Carolina Dieckmann) y la Ley 14.155/2021 sobre cibercrimen, recientemente reformada para incrementar las penas de los delitos informáticos; Colombia cuenta con la Ley 1273/2009 sobre delitos informáticos, considerada por la doctrina como uno de los marcos más completos de la región; Argentina cuenta con las Leyes 26.388 y 26.904 sobre delitos informáticos; Chile aprobó la Ley 21.459 (2022), que modernizó su marco de ciberdelitos adaptándolo al Convenio de Budapest; y Ecuador incorporó los delitos informáticos en el Código Orgánico Integral Penal (COIP) de 2014, con modificaciones posteriores (Palazzi, 2022; Sain, 2021).

Sin embargo, ninguno de estos marcos normativos aborda de manera explícita las especificidades de los crímenes algorítmicos de tercera generación: el uso de IA generativa para el fraude, los deepfakes delictivos, los ciberataques autónomos o la manipulación de sistemas de decisión automatizada. La persecución de estas conductas en la región depende de la aplicación extensiva de tipos penales genéricos (estafa, falsedad documental, acceso no autorizado, daños informáticos) cuya adecuación a las especificidades técnicas y jurídicas de los crímenes algorítmicos es parcial e insatisfactoria.

Esta situación genera un riesgo de doble vulnerabilidad: por una parte, la región puede convertirse en destino preferente de actores criminales que aprovechen la menor densidad normativa para operar desde o en el territorio latinoamericano; por otra parte, los ciudadanos de la región quedan expuestos a los daños de los crímenes algorítmicos sin disponer de las protecciones normativas que los marcos más avanzados ofrecen. La incorporación de varios países latinoamericanos al Convenio de Budapest (Brasil, Chile, Colombia, Argentina, Paraguay y Ecuador en proceso de adhesión) constituye un paso positivo en la dirección de la armonización normativa, pero debe complementarse con reformas legislativas específicas que aborden los crímenes algorítmicos de manera sustantiva.

Comparación de marcos normativos sobre crímenes algorítmicos: principales sistemas

Sistema jurídico	Instrumento base cibercrimen	Cobertura IA generativa	Deepfakes	Ataques autónomos	Vigilancia ilícita IA	Evaluación general
Unión Europea	Directiva 2013/40/UE; Convenio Budapest	Alta (Reglamento IA 2024)	Parcial (DSA; Dir. VAWG)	Parcial (NIS2)	Alta (Reglamento IA; RGPD)	Sistema más completo; sin tipos penales IA específicos
Estados Unidos	CFAA (reformado); leyes sectoriales	Media (EO 2023; sectorial)	Parcial (DEFIANCE Act 2024)	Alta (NSS 2023)	Media (sectorial; sin ley federal)	Fragmentado; jurisprudencia innovadora
China	Ley de Ciberseguridad 2017; Ley IA 2023	Alta (regulación nacional)	Alta (Reglamento deepfakes 2022)	Alta (contexto estatal)	Especial (contexto autoritario)	Cobertura amplia; preocupaciones DDHH

Sistema jurídico	Instrumento base cibercrimen	Cobertura IA generativa	Deepfakes	Ataques autónomos	Vigilancia ilícita IA	Evaluación general
España	CP arts. 197-200; 264-264b	Baja (pending reforma)	Parcial (reforma CP 2023)	Parcial (LO 7/2021)	Media (LO 3/2018; RGPD)	Marco sólido pero no actualizado a IA
Brasil	L. 12.737/2012; L. 14.155/2021	Baja	Sin tipificación específica	Parcial	Media (LGPD; marco emergente)	En transición; déficit IA generativa
Ecuador	COIP arts. 229-234	Muy baja	Sin tipificación	Parcial	Baja	Marco de primera generación; actualización urgente

Nota: Elaboración propia a partir de Clough (2021); Sieber (2020); Palazzi (2022); Comisión Europea (2024); Sain (2021).

Obstáculos estructurales para la persecución penal de los crímenes algorítmicos

El análisis comparado de los marcos normativos y de la literatura especializada permite identificar un conjunto de obstáculos estructurales que trascienden las deficiencias normativas específicas de cada ordenamiento y que afectan de manera general a la capacidad de los sistemas de justicia penal para responder eficazmente a los crímenes algorítmicos. Estos obstáculos se pueden agrupar en cuatro dimensiones.

Opacidad algorítmica y dificultades probatorias

El problema de la opacidad algorítmica, tratado en profundidad en los capítulos precedentes desde la perspectiva de la responsabilidad, adquiere una dimensión específicamente probatoria en el ámbito de los crímenes algorítmicos. La demostración de la comisión de un

crimen algorítmico requiere, con frecuencia, la acreditación de tres elementos técnicamente complejos: que un sistema de IA fue utilizado para cometer la conducta; que ese sistema operó de la manera descrita por el tipo penal; y que existe un nexo causal entre el funcionamiento del sistema y el resultado dañoso. Ninguno de estos elementos puede ser acreditado mediante los medios de prueba tradicionales sin un esfuerzo pericial de alto nivel técnico (Casey, 2019; Farid, 2022).

Las dificultades probatorias se agravan por la volatilidad de la evidencia digital (que puede ser destruida, alterada o cifrada en cuestión de segundos), por la dispersión internacional de los servidores y la infraestructura de los sistemas de IA (que requiere la activación de mecanismos de cooperación jurídica internacional cuya lentitud contrasta con la velocidad de los actores criminales), y por la escasez de peritos judiciales con las competencias técnicas necesarias para analizar sistemas de IA complejos (Europol, 2023). El Segundo Protocolo Adicional al Convenio de Budapest (2022) ha intentado agilizar los mecanismos de obtención transfronteriza de evidencia digital, pero su implementación efectiva depende de la voluntad y la capacidad técnica de los Estados parte, que es muy heterogénea.

Transnacionalidad y jurisdicción

Los crímenes algorítmicos son, por naturaleza, conductas de carácter transnacional: el actor criminal, la infraestructura tecnológica y la víctima pueden encontrarse en distintas jurisdicciones, y el sistema de IA utilizado puede estar distribuido en la nube a través de múltiples países. Esta transnacionalidad genera problemas de jurisdicción penal de extraordinaria complejidad, dado que la determinación del lugar de comisión del delito (locus delicti) es frecuentemente imposible o produce resultados concurrentes con múltiples jurisdicciones (Sieber, 2020; Clough, 2021).

Los mecanismos de cooperación jurídica internacional existentes (asistencia judicial mutua, órdenes europeas de investigación, los instrumentos del Convenio de Budapest) fueron diseñados para el crimen digital de primera y segunda generación y presentan limitaciones estructurales que los hacen inadecuados para la velocidad y complejidad de los crímenes algorítmicos. En particular, la ausencia de un mecanismo de cooperación ágil y jurídicamente robusto para la obtención de evidencia almacenada en la nube por proveedores de servicios situados en terceros países constituye uno de los déficits más graves del sistema actual, que

el Segundo Protocolo Adicional al Convenio de Budapest aborda solo parcialmente (Consejo de Europa, 2022).

Asimetría de capacidades

La persecución penal eficaz de los crímenes algorítmicos requiere capacidades técnicas, recursos humanos especializados y equipamiento forense que solo están disponibles en los sistemas de justicia y en los cuerpos de seguridad de los países más desarrollados. En la mayor parte de los países en vías de desarrollo, incluida una significativa proporción de los países latinoamericanos, la brecha de capacidades entre los actores criminales tecnológicamente sofisticados y las agencias de persecución penal es de primera magnitud (Interpol, 2023; Sain, 2021). Esta asimetría no solo reduce la eficacia de la persecución, sino que tiene un efecto de desplazamiento geográfico: los actores criminales tienden a operar desde jurisdicciones con menor capacidad de detección y respuesta.

El fortalecimiento de las capacidades nacionales en materia de ciberseguridad y persecución del cibercrimen es, por tanto, una condición necesaria para la efectividad de cualquier reforma normativa. En este sentido, los programas de capacitación y asistencia técnica de organismos internacionales como Interpol, Europol, UNODC y el Consejo de Europa constituyen un complemento indispensable a las reformas legislativas, y su financiación adecuada debería ser considerada una prioridad de política criminal internacional (Interpol, 2023; UNODC, 2023).

Velocidad de mutación tecnológica

El cuarto obstáculo estructural es la velocidad de mutación de las herramientas y modalidades de los crímenes algorítmicos, que supera con creces el ritmo de los ciclos legislativos ordinarios. En el tiempo que tarda un parlamento en elaborar, debatir y aprobar una ley específica sobre deepfakes o sobre ataques autónomos, las herramientas tecnológicas disponibles para los actores criminales habrán experimentado varias generaciones de mejora. Esta realidad impone la adopción de estrategias normativas de mayor flexibilidad que las tipificaciones penales rígidas tradicionales: normas de habilitación que permitan la incorporación ágil de nuevas modalidades, cláusulas abiertas de remisión a estándares técnicos actualizables, y procedimientos parlamentarios de urgencia para la respuesta a amenazas emergentes (Hildebrandt, 2020; Floridi et al., 2022).

DISCUSIÓN Y CONCLUSIONES

Discusión

Los resultados presentados en el apartado anterior permiten formular una serie de reflexiones críticas que sitúan los hallazgos tipológicos y comparados en el contexto más amplio del debate sobre la respuesta del derecho penal al fenómeno de los crímenes algorítmicos.

Una primera reflexión concierne a la adecuación de las estrategias normativas disponibles para enfrentar los crímenes algorítmicos. El debate doctrinal presenta dos posiciones principales. La primera, que se puede denominar posición adaptacionista, sostiene que los sistemas jurídicos cuentan con recursos dogmáticos suficientes para dar respuesta a los crímenes algorítmicos mediante la aplicación extensiva y la reinterpretación de los tipos penales existentes, complementada con reformas técnicas específicas cuando sea estrictamente necesario (Sieber, 2020; Clough, 2021). Esta posición tiene la ventaja de evitar la inflación legislativa y de preservar la coherencia sistemática del derecho penal, pero corre el riesgo de forzar los tipos penales más allá de los límites que el principio de legalidad (nullum crimen sine lege) permite.

La segunda posición, que se puede denominar posición reformista estructural, sostiene que los crímenes algorítmicos son fenomenológicamente tan distintos de los delitos informáticos clásicos que requieren nuevos tipos penales específicos, nuevas reglas de imputación y nuevos instrumentos procesales diseñados para sus características particulares (Brundage et al., 2021; Ferrara, 2023). Esta posición tiene la ventaja de la especificidad y la adecuación técnica, pero corre los riesgos de la hiperespecialización normativa y de la obsolescencia acelerada de las tipificaciones en un campo tecnológico en rápida evolución.

Los hallazgos del presente capítulo sugieren que ninguna de estas posiciones es suficiente por sí sola, y que la respuesta más adecuada es una estrategia mixta que combine la actualización interpretativa de los tipos existentes (posición adaptacionista) con la tipificación específica de las modalidades más lesivas y más claramente diferenciadas de los crímenes algorítmicos, como los deepfakes delictivos, los ataques autónomos contra infraestructuras críticas y la manipulación de sistemas de decisión de interés público (posición reformista).

Una segunda reflexión concierne a la tensión entre eficacia punitiva y garantismo en el ámbito de los crímenes algorítmicos. La eficacia en la persecución de estas conductas exige, en

muchos casos, medidas que entran en tensión con derechos fundamentales: el cifrado de comunicaciones (que protege la privacidad pero dificulta la obtención de evidencia), el anonimato en Internet (que protege la libertad de expresión pero facilita la impunidad), la velocidad de los procedimientos de cooperación internacional (que puede requerir mecanismos de urgencia que limiten las garantías procesales ordinarias). Estas tensiones no son nuevas en el derecho penal, pero adquieren una especial intensidad en el ámbito de los crímenes algorítmicos dada la escala potencial de los daños y la sofisticación de los actores. La resolución de estas tensiones requiere, como han señalado Hildebrandt (2020) y Pasquale (2020), un debate democrático fundamentado en los principios constitucionales del Estado de Derecho, y no puede ser dejada a la discrecionalidad de los actores tecnológicos o de las agencias de seguridad.

Una tercera reflexión versa sobre las implicaciones de los hallazgos para el contexto latinoamericano, al que los capítulos precedentes de esta obra han prestado atención específica. La constatación de que los países de la región cuentan, en su mayoría, con marcos normativos de primera o segunda generación que resultan insuficientes para enfrentar los crímenes algorítmicos de tercera generación, unida a la constatación de que la región está experimentando un crecimiento acelerado del despliegue de sistemas de IA en ámbitos críticos como la seguridad pública, la justicia y los servicios financieros, configura una situación de vulnerabilidad normativa que debería impulsar reformas legislativas urgentes. La adhesión al Convenio de Budapest y la adopción de legislaciones específicas de cibercrimen actualizadas, complementadas por programas de fortalecimiento de capacidades técnicas y de cooperación regional, son los tres pilares de una estrategia normativa regionalmente adecuada.

En relación con los trabajos previos, los hallazgos del presente capítulo son consistentes con las conclusiones de Clough (2021) sobre el desfase normativo del derecho comparado ante las nuevas modalidades de cibercrimen, y con las observaciones de Brundage et al. (2021) sobre el potencial de la IA para transformar el paisaje de la amenaza criminal. Al mismo tiempo, el presente análisis complementa y actualiza esos trabajos incorporando las implicaciones del Reglamento de IA europeo de 2024, la Convención de la ONU sobre Ciberdelincuencia de 2024 y los desarrollos jurisprudenciales más recientes en materia de crímenes algorítmicos.

Una consideración adicional concierne a la dimensión internacional y a la necesidad de convergencia normativa global. La naturaleza transnacional de los crímenes algorítmicos hace que la eficacia de cualquier reforma nacional esté condicionada por el nivel de armonización normativa internacional. La Convención de las Naciones Unidas sobre Ciberdelincuencia, aprobada en 2024 tras un prolongado proceso de negociación, constituye un avance significativo en esta dirección, aunque su texto ha sido objeto de críticas por parte de organizaciones de derechos humanos y de la sociedad civil, que señalan que algunas de sus disposiciones podrían ser utilizadas por regímenes autoritarios para criminalizar la actividad legítima de periodistas, activistas y investigadores en seguridad informática (EFF, 2023; Access Now, 2023). Esta tensión ilustra el dilema fundamental que toda iniciativa de armonización global del derecho del cibercrimen enfrenta: la necesidad de cooperación internacional eficaz choca con la diversidad de los sistemas políticos y de los estándares de derechos humanos entre los Estados que deben cooperar.

Conclusiones

Del análisis efectuado en el presente capítulo es posible extraer las siguientes conclusiones, formuladas con la cautela epistemológica que el carácter documental de la investigación y el estado de evolución del campo imponen.

Los crímenes algorítmicos constituyen una categoría delictiva emergente, cualitativa y cuantitativamente distinta de los delitos informáticos clásicos, que se caracteriza por el papel funcionalmente determinante de los sistemas de IA en la comisión de las conductas delictivas. Las principales modalidades identificadas son: el fraude algorítmico, los deepfakes delictivos, los ciberataques autónomos, la manipulación de sistemas de decisión automatizada, los crímenes de datos y la desinformación algorítmica. Cada una de estas modalidades presenta especificidades dogmáticas y probatorias que las tipologías penales clásicas no capturan de manera satisfactoria.

Los marcos normativos vigentes en los principales sistemas jurídicos comparados presentan un desfase significativo respecto a la realidad de los crímenes algorítmicos de tercera generación. El sistema europeo es el más desarrollado, gracias a la combinación del Reglamento de IA, la Directiva NIS2 y el DSA, pero carece de tipos penales específicos para las conductas más lesivas. Los sistemas de América Latina presentan los déficits más

pronunciados, con marcos de primera o segunda generación que resultan manifiestamente insuficientes ante las amenazas actuales.

La persecución penal eficaz de los crímenes algorítmicos enfrenta cuatro obstáculos estructurales que trascienden las deficiencias normativas específicas: la opacidad algorítmica y sus implicaciones probatorias; la transnacionalidad de las conductas y los límites de los mecanismos de cooperación internacional; la asimetría de capacidades técnicas entre actores criminales y agencias de persecución; y la velocidad de mutación de las herramientas tecnológicas disponibles para los actores delictivos.

La estrategia normativa más adecuada para hacer frente a los crímenes algorítmicos no es la de la hiperespecialización legislativa ni la de la mera extensión interpretativa de los tipos existentes, sino una estrategia mixta que combine la actualización del contenido de los tipos penales existentes mediante la incorporación de estándares técnicos de referencia, con la tipificación específica de las modalidades más lesivas y más claramente diferenciadas, y con la adopción de mecanismos procesales adaptados a las características probatorias de los crímenes algorítmicos.

La dimensión internacional es insoslayable en el tratamiento normativo de los crímenes algorítmicos. La Convención de las Naciones Unidas sobre Ciberdelincuencia (2024) y el fortalecimiento del régimen del Convenio de Budapest constituyen los marcos de referencia obligados, pero deben ser complementados con mecanismos regionales de cooperación, especialmente en América Latina, y con programas robustos de fortalecimiento de capacidades técnicas en los países con menor desarrollo institucional.

El presente capítulo identifica como prioritarias para la investigación futura las siguientes líneas de trabajo: el análisis empírico-comparado de la jurisprudencia penal en casos de crímenes algorítmicos en distintas jurisdicciones; el desarrollo de propuestas específicas de tipificación penal para los deepfakes delictivos y los ataques autónomos contra infraestructuras críticas, adaptadas al contexto jurídico latinoamericano; y la evaluación del impacto de la Convención de la ONU sobre Ciberdelincuencia de 2024 sobre los estándares de derechos humanos en los Estados que la adopten.

En conclusión, los delitos informáticos y crímenes algorítmicos representan uno de los frentes más urgentes e insuficientemente atendidos del derecho penal contemporáneo. Su adecuado

tratamiento normativo es condición necesaria no solo para la protección de las víctimas individuales, sino para la preservación de la integridad de los sistemas democráticos, económicos y de justicia que los sistemas de IA amenazan en caso de ser instrumentalizados con fines delictivos. La elaboración de respuestas dogmáticas, normativas y de cooperación internacional a la altura de este desafío es una tarea que el presente capítulo ha procurado adelantar, con la conciencia de que el ritmo de los cambios tecnológicos exigirá continuar este esfuerzo de manera sistemática e interdisciplinaria.

CAPITULO IV

Prueba Digital y Evidencia Algorítmica en Juicio Penal

Prueba Digital y Evidencia Algorítmica en Juicio Penal

YUNIQUER RICARDO AVENDAÑO VERA[4]

RESUMEN

El presente capítulo analiza la problemática de la prueba digital y la evidencia algorítmica en el juicio penal como una de las dimensiones más críticas y menos consolidadas de la intersección entre la inteligencia artificial y el sistema de justicia. A partir de una revisión bibliográfica sistemática de fuentes especializadas indexadas en Scopus, Web of Science y HeinOnline, correspondientes al período 2018-2025, y del análisis comparado de marcos normativos y jurisprudencia de la Unión Europea, Estados Unidos, América Latina y el sistema del Convenio de Budapest, se examinan los fundamentos técnicos y jurídicos de la evidencia digital, los estándares de admisibilidad probatoria aplicables a las conclusiones algorítmicas, las garantías del proceso penal en riesgo de erosión y los modelos de peritaje forense digital emergentes. Los resultados revelan una brecha normativa y jurisprudencial de primera magnitud entre la sofisticación técnica de los sistemas generadores de evidencia y la capacidad del derecho procesal penal para someterlos a control de contradicción, motivación y tutela efectiva. Se concluye que la superación de esta brecha requiere la convergencia de reformas procesales, estándares técnicos de certificación forense y un estatuto probatorio específico para la evidencia algorítmica que preserve las garantías del Estado de Derecho.

Palabras clave: *prueba digital; evidencia algorítmica; juicio penal; cadena de custodia; admisibilidad probatoria; pericia forense digital; inteligencia artificial; garantías procesales; derecho a la prueba; opacidad algorítmica.*

[4] Abogado de los Juzgados y Tribunales de la República del Ecuador, Universidad Guayaquil, Tecnólogo Superior en Criminalística, Instituto Tecnológico Universitario Argos, Consultor Externo, correo electrónico: yuniquer-97@hotmail.com.

ABSTRACT

This chapter analyzes the problem of digital evidence and algorithmic evidence in criminal proceedings as one of the most critical and least consolidated dimensions of the intersection between artificial intelligence and the justice system. Based on a systematic literature review of specialized sources indexed in Scopus, Web of Science and HeinOnline for the period 2018–2025, and a comparative analysis of regulatory frameworks and jurisprudence from the European Union, the United States, Latin America, and the Budapest Convention system, the chapter examines the technical and legal foundations of digital evidence, the admissibility standards applicable to algorithmic conclusions, the criminal procedural guarantees at risk of erosion, and the emerging models of digital forensic expertise. The findings reveal a significant normative and jurisprudential gap between the technical sophistication of evidence-generating systems and the capacity of criminal procedural law to subject them to adversarial scrutiny, motivation requirements, and effective judicial protection. The chapter concludes that closing this gap requires the convergence of procedural reforms, technical standards for forensic certification, and a specific evidentiary statute for algorithmic evidence that preserves the guarantees of the rule of law.

Keywords: *digital evidence; algorithmic evidence; criminal proceedings; chain of custody; evidentiary admissibility; digital forensic expertise; artificial intelligence; procedural guarantees; right to evidence; algorithmic opacity.*

INTRODUCCIÓN

El juicio penal ha sido concebido históricamente como un espacio epistémico de alta exigencia: el Estado, al ejercer su poder punitivo, está obligado a demostrar la culpabilidad del acusado mediante un estándar de prueba que, en los sistemas de herencia angloamericana, se formula como la ausencia de duda razonable (beyond reasonable doubt), y que en los sistemas continentales se expresa en la exigencia de plena convicción judicial sustentada en prueba lícitamente obtenida y practicada con contradicción (Miranda Estrampes, 2022). Este estándar elevado no es un capricho procesal, sino la expresión normativa del principio de presunción de inocencia: si el Estado no puede demostrar la culpabilidad más allá de toda duda razonable, el ciudadano debe ser absuelto, aun cuando ello implique que un culpable quede en libertad. La robustez del sistema de justicia penal depende, en última instancia, de la calidad epistemológica de la prueba sobre la que se funda la condena.

La irrupción de la inteligencia artificial como fuente, procesadora y analizadora de evidencia en el proceso penal introduce perturbaciones de primera magnitud en este edificio epistemológico. Los sistemas algorítmicos generan, seleccionan y procesan cantidades ingentes de datos que pueden resultar relevantes para la determinación de los hechos en juicio: registros de geolocalización de teléfonos móviles, patrones de comportamiento inferidos del tráfico de internet, análisis de reconocimiento facial en grabaciones de videovigilancia, predicciones de reincidencia elaboradas por herramientas de evaluación de riesgo, transcripciones generadas por sistemas de procesamiento del lenguaje natural, o conclusiones periciales producidas por sistemas de análisis forense automatizado (Casey, 2019; Kerr, 2020; Raeini et al., 2023). El punto crítico es que, en muchos de estos supuestos, la lógica interna del proceso mediante el cual el sistema algorítmico llegó a su conclusión no puede ser examinada, comprendida ni cuestionada mediante los mecanismos procesales ordinarios de contradicción, lo que introduce un déficit estructural de garantías que compromete la legitimidad epistémica del juicio penal.

La dimensión del problema se hace patente al constatar que los sistemas de inteligencia artificial ya son utilizados de manera rutinaria en diversas etapas del proceso penal en numerosos países. En los Estados Unidos, los sistemas de evaluación de riesgo algorítmico (risk assessment tools) como COMPAS, PSA o LSI-R se emplean en más de dos tercios de los estados federales para fundamentar decisiones judiciales sobre prisión preventiva, libertad

condicional y sentencia (Angwin et al., 2016; Dressel & Farid, 2022). En el Reino Unido y los Países Bajos, los sistemas de policing predictivo basan sus recomendaciones de patrullaje y vigilancia en análisis algorítmicos de patrones delictivos históricos. En la Unión Europea, los sistemas de reconocimiento facial son utilizados por las fuerzas de seguridad de varios Estados miembros en el marco de investigaciones criminales, y sus conclusiones son incorporadas como elemento probatorio en los procesos penales (Europol, 2023; Garvie et al., 2022). En América Latina, el uso de evidencia digital de todo tipo (mensajes de texto, comunicaciones de aplicaciones cifradas, metadatos de dispositivos móviles, imágenes de videovigilancia procesadas algorítmicamente) se ha convertido en un elemento central de la estrategia probatoria de las fiscalías en delitos contra la vida, la libertad sexual, el patrimonio y el tráfico de estupefacientes (Palazzi, 2022; Ortega & Salinas, 2023).

La relevancia académica y práctica del estudio de la prueba digital y la evidencia algorítmica en el juicio penal se asienta sobre tres grandes ejes de fundamentación. En primer lugar, la admisibilidad de la evidencia digital plantea problemas conceptuales y procesales que los marcos normativos vigentes, diseñados para un paradigma probatorio de prueba física y testimonial, no están en condiciones de resolver de manera satisfactoria. Las reglas clásicas de autenticación, relevancia y fiabilidad requieren una reinterpretación profunda cuando se aplican a conclusiones generadas por sistemas que operan mediante procesos de aprendizaje automático cuyo funcionamiento interno no puede ser explicado de manera comprensible para las partes ni para el tribunal (Doshi-Velez & Kim, 2021; Wachter et al., 2020). En segundo lugar, la cadena de custodia de la evidencia digital presenta desafíos técnicos específicos derivados de la volatilidad, la duplicabilidad y la mutabilidad de los datos digitales, que exigen protocolos forenses especializados cuya ausencia o deficiencia puede comprometer irreparablemente el valor probatorio de la evidencia (Casey, 2019; Karie & Venter, 2020). En tercer lugar, el principio de contradicción (cornerstone del proceso penal acusatorio) se ve seriamente amenazado cuando la defensa no tiene acceso a la información técnica necesaria para cuestionar la fiabilidad y la exactitud de las conclusiones algorítmicas, bien porque el sistema es propietario y está protegido por secreto comercial, bien porque su complejidad técnica excede las posibilidades de comprensión de cualquier perito no especialista (Pasquale, 2020; Ferguson, 2021).

El debate sobre estos problemas ha producido en los últimos años una literatura académica de notable riqueza y complejidad. Los trabajos fundacionales de Casey (2019) sobre investigación forense digital sentaron las bases técnicas y metodológicas del análisis. Las aportaciones de Kerr (2020) sobre el derecho de la evidencia digital y la Cuarta Enmienda en el contexto estadounidense, de Hildebrandt (2020) sobre el rule of law en la era de los sistemas inteligentes, de Završnik (2021) sobre la justicia penal algorítmica, y de Ferguson (2021) sobre el uso policial de la inteligencia artificial, han contribuido a definir los términos del debate desde distintas perspectivas disciplinares. Más recientemente, los trabajos de Raeini et al. (2023) sobre los estándares de admisibilidad de la evidencia forense digital, de Maras & Alexandrou (2023) sobre la autenticación de la evidencia mediada por IA, y de Cheng et al. (2024) sobre la explicabilidad algorítmica en el proceso judicial, han avanzado propuestas de mayor especificidad normativa. En el ámbito latinoamericano, las contribuciones de Palazzi (2022) sobre el derecho penal informático, de Ortega y Salinas (2023) sobre la gobernanza de la IA en América Latina, y de Miranda Estrampes (2022) sobre la prueba digital en el proceso penal iberoamericano, constituyen referencias indispensables para contextualizar el análisis en el marco del derecho continental hispanohablante.

Desde una perspectiva de derecho comparado, los ordenamientos jurídicos han dado respuestas muy heterogéneas al problema de la evidencia algorítmica. En los Estados Unidos, la jurisprudencia ha evolucionado desde la aplicación del estándar Frye (1923) de general acceptance científica hasta el test Daubert (1993), que impone al juez un rol de gatekeeper activo en la evaluación de la fiabilidad metodológica de la prueba pericial. Algunos tribunales han comenzado a aplicar estos estándares a la evidencia algorítmica, con resultados aún inconsistentes (Cheng et al., 2024; Kerr, 2020). En la Unión Europea, el Reglamento de Inteligencia Artificial de 2024 (UE 2024/1689) clasifica como de alto riesgo los sistemas de IA utilizados en el ámbito de la justicia penal y establece requisitos de transparencia y explicabilidad, pero no contiene disposiciones procesales penales específicas sobre la admisibilidad de la evidencia generada por dichos sistemas, materia que continúa siendo competencia de los Estados miembros (Comisión Europea, 2024). En América Latina, la regulación específica de la evidencia digital en el proceso penal es todavía escasa y fragmentaria, con algunas excepciones notables en la legislación procesal de Colombia, Chile y Argentina (Palazzi, 2022; Miranda Estrampes, 2022).

El propósito del presente capítulo es, en consecuencia, cuádruple. En primer lugar, establecer los fundamentos técnicos y jurídicos de la prueba digital y la evidencia algorítmica, definiendo sus categorías constitutivas y sus especificidades respecto de la prueba tradicional. En segundo lugar, analizar los estándares de admisibilidad probatoria aplicables a la evidencia algorítmica desde una perspectiva comparada. En tercer lugar, identificar las garantías procesales penales en riesgo de erosión y examinar los mecanismos de protección disponibles o necesarios. En cuarto lugar, evaluar los modelos de peritaje forense digital emergentes y proponer criterios normativos para un estatuto probatorio de la evidencia algorítmica compatible con las garantías del Estado de Derecho. El capítulo se inscribe en la continuidad temática establecida por los capítulos precedentes de la presente obra, que han abordado los fundamentos conceptuales del derecho penal ante la IA (Capítulo I), la responsabilidad penal en sistemas autónomos (Capítulo II) y la tipología de los delitos informáticos y crímenes algorítmicos (Capítulo III). La cuestión de la prueba y la evidencia constituye el eslabón lógicamente siguiente en esta cadena analítica, pues sin una teoría adecuada de la evidencia algorítmica, los problemas de tipificación e imputación abordados en los capítulos anteriores carecen de solución procesal práctica.

MATERIAL Y MÉTODOS

Diseño de Investigación

El presente capítulo se inscribe en el marco de una investigación de naturaleza teórico-documental y de derecho comparado, coherente con la estrategia metodológica adoptada en los capítulos precedentes de la obra. El diseño combina la revisión bibliográfica sistemática con el análisis dogmático-jurídico y comparado de fuentes secundarias especializadas. Esta elección metodológica resulta la más adecuada para el objeto de estudio por cuanto la problemática de la prueba digital y la evidencia algorítmica es fundamentalmente una cuestión de construcción conceptual, interpretación jurídica y análisis normativo comparado, que no requiere ni se beneficiaría de un trabajo de campo cuantitativo en el estado actual del desarrollo del campo. El diseño se enmarca en la tradición metodológica del derecho procesal comparado, siguiendo los estándares de rigor analítico propios de la investigación jurídica académica (Siems, 2018; Zweigert & Kötz, 2020).

Fuentes y Corpus Documental

La construcción del corpus documental se realizó mediante búsquedas sistemáticas en las bases de datos Scopus, Web of Science, HeinOnline, Westlaw International y SSRN durante el período comprendido entre julio de 2024 y marzo de 2025. Los descriptores de búsqueda empleados incluyeron los términos: "digital evidence criminal proceedings", "algorithmic evidence admissibility", "artificial intelligence criminal trial", "forensic digital evidence AI", "chain of custody digital evidence", "explainability algorithmic evidence", "predictive policing evidence", "facial recognition criminal evidence", "machine learning evidence standards", así como sus equivalentes en español y francés. Se priorizaron artículos publicados en revistas indexadas en los cuartiles Q1 y Q2 según el Journal Citation Reports, así como monografías y capítulos de obras colectivas de reconocido prestigio en el ámbito del derecho procesal penal, el derecho probatorio y la informática jurídica.

Se incorporaron asimismo al corpus los siguientes instrumentos normativos de primer orden: el Convenio de Budapest sobre Ciberdelincuencia (2001) y su Segundo Protocolo Adicional (2022); el Reglamento de Inteligencia Artificial de la Unión Europea (UE 2024/1689); las Directrices del Consejo de Europa sobre uso de IA en el sistema judicial (CM/Rec(2020)1); los Principios de la OCDE sobre IA (2023); el Reglamento General de Protección de Datos (UE 2016/679); la jurisprudencia relevante del Tribunal Europeo de Derechos Humanos (TEDH), el Tribunal de Justicia de la Unión Europea (TJUE) y la Corte Suprema de los Estados Unidos; así como las legislaciones procesales penales de Alemania, España, Francia, Estados Unidos, Colombia, Argentina, Chile y Ecuador en materia de prueba digital. El corpus final procesado comprende 198 fuentes documentales, de las cuales se seleccionaron 46 como referencias directamente citadas en el texto, atendiendo a criterios de pertinencia temática, rigor metodológico, actualidad y representatividad geográfica y lingüística.

Procesamiento y Análisis

El análisis de los materiales seleccionados se llevó a cabo mediante la técnica del análisis documental cualitativo, que comprende las fases de lectura comprensiva, identificación de argumentos centrales, construcción de categorías analíticas y síntesis interpretativa crítica. Se adoptó un enfoque dogmático-jurídico para el tratamiento de las fuentes de derecho positivo y doctrinal, y un enfoque hermenéutico para la interpretación de los documentos de política

pública, los estándares técnicos y los informes institucionales. La triangulación de perspectivas disciplinares (derecho procesal penal, derecho probatorio comparado, ciencia forense digital, ética de la tecnología e inteligencia artificial) se empleó como estrategia de validación interna del análisis, con el objetivo de asegurar la coherencia y completitud de las categorías construidas.

El método comparativo se empleó en su modalidad funcional (Zweigert & Kötz, 2020), partiendo de los problemas jurídicos que la evidencia algorítmica plantea de manera común en distintos ordenamientos, para examinar después las soluciones que cada sistema jurídico ofrece. Esta aproximación permite identificar convergencias y divergencias significativas entre los modelos normativos y extraer enseñanzas de política criminal y procesal aplicables a contextos diversos.

Limitaciones Metodológicas

Entre las principales limitaciones del diseño adoptado conviene señalar las siguientes. En primer lugar, la acelerada evolución del campo tecnológico y normativo implica que algunos de los desarrollos jurisprudenciales o legislativos descritos pueden haber experimentado modificaciones en el momento de la publicación del presente capítulo. En segundo lugar, la prevalencia de literatura académica en inglés y en el contexto angloamericano introduce un sesgo de selección geográfico que se ha intentado compensar con la incorporación sistemática de fuentes doctrinales de la tradición continental europea e iberoamericana. En tercer lugar, la metodología documental no permite validar empíricamente las propuestas normativas examinadas mediante estudios de impacto real sobre la práctica judicial, limitación que deberá ser subsanada en investigaciones futuras de naturaleza empírico-comparada basadas en análisis de expedientes judiciales y encuestas a operadores jurídicos.

RESULTADOS

Fundamentos Técnicos y Jurídicos de la Prueba Digital y la Evidencia Algorítmica

Concepto y categorías de la prueba digital

La revisión de la literatura permite identificar una distinción conceptual fundamental que la doctrina procesal no siempre traza con claridad suficiente: la diferencia entre prueba digital y evidencia algorítmica. La prueba digital (digital evidence) es un concepto más amplio que

engloba cualquier información almacenada o transmitida en formato digital que pueda resultar relevante para la determinación de los hechos en un proceso penal. Casey (2019) la define, en la formulación más ampliamente aceptada, como cualquier dato que puede establecer que un delito ha tenido lugar o puede proveer un vínculo entre un delito y su víctima, o entre un delito y su autor. Esta categoría abarca desde los mensajes de correo electrónico y las comunicaciones de mensajería instantánea hasta los registros de acceso a sistemas informáticos, los metadatos de archivos digitales, los datos de geolocalización de dispositivos móviles, las grabaciones de videovigilancia y los registros de actividad de vehículos con sistemas telemáticos.

La evidencia algorítmica constituye una subcategoría específica y más compleja de la prueba digital, caracterizada por el hecho de que su contenido probatorio no es directamente perceptible sino que es generado, inferido o seleccionado mediante procesos algorítmicos de análisis, clasificación o predicción. Se trata, en suma, de las conclusiones producidas por un sistema de inteligencia artificial a partir del procesamiento de datos digitales: la identificación de una persona como el probable autor de un delito por un sistema de reconocimiento facial; la predicción de reincidencia delictiva por un sistema de evaluación de riesgo actuarial; la determinación de la autoría de un texto mediante análisis estilométrico automatizado; o la reconstrucción de la cronología de eventos mediante el análisis algorítmico de metadatos (Raeini et al., 2023; Maras & Alexandrou, 2023).

Desde una perspectiva jurídico-procesal, la evidencia algorítmica introduce complejidades específicas que no presentan las formas más simples de prueba digital. Su producción está mediada por procesos técnicos que el tribunal y las partes no están en condiciones de examinar directamente. Su fiabilidad depende de variables técnicas (calidad de los datos de entrenamiento, arquitectura del modelo, parámetros de funcionamiento) que requieren conocimiento especializado para ser evaluadas. Y su alcance probatorio es frecuentemente estadístico o probabilístico (el sistema asigna una probabilidad de que el acusado sea el autor, o de que reincida) antes que categórico, lo que plantea el problema de la traducción de la incertidumbre estadística en la convicción judicial que el estándar de prueba penal requiere (Doshi-Velez & Kim, 2021; Cheng et al., 2024).

Características técnicas diferenciales de la evidencia digital

La doctrina forense especializada, con Casey (2019) y Karie & Venter (2020) como referentes principales, ha identificado un conjunto de características técnicas que distinguen la evidencia digital de la evidencia física tradicional y que condicionan su tratamiento procesal. La primera característica es la volatilidad: la evidencia digital puede desaparecer, alterarse o corromperse de manera accidental o deliberada con una facilidad y una velocidad que no tiene parangón en la evidencia física. Los datos almacenados en memoria RAM desaparecen al desconectar el dispositivo; los logs de conexión de los proveedores de servicios de internet son borrados rutinariamente tras períodos breves; los metadatos de los archivos pueden ser modificados sin dejar rastro en los sistemas de archivos más comunes. Esta volatilidad impone la necesidad de protocolos de captura y preservación inmediata que, si no se siguen adecuadamente, pueden contaminar o destruir la evidencia.

La segunda característica es la duplicabilidad perfecta: a diferencia de la evidencia física, la evidencia digital puede ser copiada de manera exacta e ilimitada sin degradación ni alteración del original. Esta característica tiene implicaciones procesales ambivalentes: por una parte, facilita la conservación de copias de seguridad y la distribución de la evidencia entre las partes; por otra, hace imposible distinguir el original de la copia mediante métodos organolépticos, lo que exige el uso de mecanismos criptográficos de verificación de integridad (funciones hash) para acreditar la autenticidad de la cadena de custodia (Karie & Venter, 2020; Reyes & Lutteral, 2021).

La tercera característica es la dispersión geográfica y jurisdiccional: los datos digitales relevantes para un proceso penal pueden estar almacenados en servidores ubicados en múltiples países, gestionados por empresas bajo distintas jurisdicciones y sometidos a regulaciones divergentes sobre privacidad y acceso de las autoridades. Esta dispersión genera fricciones de cooperación internacional que pueden hacer prácticamente inaccesibles datos cruciales para la investigación, o que pueden requerir años de gestión para su obtención por vías de asistencia judicial mutua (Clough, 2021; Consejo de Europa, 2022).

La cuarta característica, específica de la evidencia algorítmica, es la opacidad o caja negra: muchos sistemas de IA, especialmente aquellos basados en redes neuronales profundas (deep learning), toman sus decisiones mediante procesos internos que no pueden ser explicitados

en términos comprensibles para los no especialistas. El propio desarrollador del sistema puede ser incapaz de explicar por qué el sistema llegó a una conclusión específica en un caso concreto. Esta opacidad hace imposible la reproducción del razonamiento probatorio en términos que el proceso penal exige, y plantea una amenaza directa al derecho de defensa y al principio de motivación de las resoluciones judiciales (Pasquale, 2020; Wachter et al., 2020; Arrieta et al., 2020).

Características técnicas de la evidencia digital y sus implicaciones procesales

Característica	Descripción técnica	Implicación procesal penal	Referencia clave
Volatilidad	Posibilidad de alteración o destrucción accidental o deliberada en tiempo mínimo	Exigencia de protocolos de captura y preservación inmediata; riesgo de pérdida irreparable de prueba	Casey (2019); Karie & Venter (2020)
Duplicabilidad perfecta	Copiado exacto e ilimitado sin degradación del original	Necesidad de funciones hash para verificación de integridad; imposibilidad de distinción organoléptica original/copia	Reyes & Lutteral (2021)
Dispersión jurisdiccional	Almacenamiento en servidores distribuidos en múltiples países y jurisdicciones	Fricciones de cooperación internacional; riesgo de inaccesibilidad de datos; aplicación concurrente de marcos legales divergentes	Clough (2021); Consejo de Europa (2022)

Característica	Descripción técnica	Implicación procesal penal	Referencia clave
Opacidad algorítmica	Imposibilidad de explicar el proceso decisorio interno de sistemas de IA complejos	Amenaza al derecho de defensa, al principio de contradicción y al deber de motivación judicial	Wachter et al. (2020); Pasquale (2020)
Trazabilidad técnica	Capacidad de los datos digitales de registrar metadatos sobre su origen, modificación y acceso	Oportunidad para la reconstrucción forense de cronologías y autoría; exige pericia especializada	Casey (2019); Raeini et al. (2023)

Nota: Elaboración propia a partir de Casey (2019), Karie & Venter (2020), Clough (2021), Wachter et al. (2020), Reyes & Lutteral (2021) y Raeini et al. (2023).

Estándares de Admisibilidad de la Evidencia Algorítmica: Análisis Comparado

El sistema norteamericano: de Frye a Daubert y su proyección sobre la evidencia algorítmica

El ordenamiento jurídico estadounidense ha desarrollado el marco más elaborado de estándares de admisibilidad de la prueba científica y técnica, cuya aplicación a la evidencia algorítmica está comenzando a ser analizada por la jurisprudencia y la doctrina. El estándar Frye, formulado en Frye v. United States (293 F. 1013, D.C. Cir. 1923), establecía que la prueba científica es admisible cuando el método en que se funda ha alcanzado aceptación general en la comunidad científica relevante. Aunque este estándar fue reemplazado en el nivel federal por el test Daubert (Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579, 1993), continúa vigente en varios estados.

El test Daubert, incorporado a la Regla 702 de las Reglas Federales de Evidencia, impone al juez el rol de gatekeeper epistémico, exigiendo que la prueba pericial sea: (1) basada en

suficientes hechos y datos; (2) producto de principios y métodos fiables; y (3) aplicada de manera fiable a los hechos del caso. La jurisprudencia posterior (Kumho Tire Co. v. Carmichael, 526 U.S. 137, 1999) extendió este estándar a toda prueba pericial de naturaleza técnica o especializada, no solo científica. La aplicación de estos estándares a la evidencia algorítmica plantea dificultades específicas: el criterio de «principios y métodos fiables» exige un nivel de transparencia sobre el funcionamiento del sistema que muchos algoritmos de inteligencia artificial, especialmente los basados en deep learning, no pueden proporcionar; el criterio de «aceptación general» puede resultar difícil de acreditar en campos tecnológicos que evolucionan con mayor rapidez que la producción científica revisada por pares (Cheng et al., 2024; Kerr, 2020).

Un hito jurisprudencial de especial relevancia en este ámbito es el asunto State v. Loomis (881 N.W.2d 749, Wis. 2016), en el que el Tribunal Supremo de Wisconsin avaló el uso del sistema COMPAS en la fase de sentencia, pero estableció que el juez no puede fundamentar la condena exclusivamente en la valoración de riesgo algorítmica, que la defensa debe tener acceso a información suficiente sobre el método para poder cuestionarlo, y que las valoraciones de riesgo basadas en datos grupales no pueden utilizarse para determinar la culpabilidad individual del acusado. Esta decisión ha sido ampliamente analizada por la doctrina como una primera respuesta judicial tentativa a los problemas de la evidencia algorítmica, que plantea más interrogantes de los que resuelve (Dressel & Farid, 2022; Završnik, 2021).

El sistema europeo: garantías del CEDH y el Reglamento de IA

El marco europeo de protección de garantías procesales penales se articula principalmente en torno al artículo 6 del Convenio Europeo de Derechos Humanos (CEDH), que consagra el derecho a un proceso equitativo (fair trial), y a la jurisprudencia del Tribunal Europeo de Derechos Humanos (TEDH) sobre su contenido. El TEDH ha establecido que el derecho a un proceso equitativo incluye el derecho a conocer y contradecir toda la evidencia presentada en contra del acusado, el derecho a cuestionar la fiabilidad de las pruebas periciales, y la obligación del tribunal de motivar suficientemente sus resoluciones con referencia a las pruebas en que se fundamentan (TEDH, Rowe y Davis c. Reino Unido, 2000; TEDH, Al-Khawaja y Tahery c. Reino Unido, 2011). La aplicación de estos principios a la evidencia algorítmica exige que el acusado tenga acceso a información suficiente sobre el método algorítmico

utilizado para poder cuestionarlo efectivamente, y que el tribunal sea capaz de motivar su decisión con referencia a los elementos de fiabilidad del sistema (Hildebrandt, 2020; Završnik, 2021).

La reciente jurisprudencia del TJUE ha reforzado estas exigencias en el contexto de la protección de datos. En el asunto C-817/19 (Ligue des droits humains v. Conseil des ministres, 2022), el TJUE declaró que el tratamiento automatizado de datos personales a gran escala con fines de seguridad pública debe estar sujeto a garantías efectivas de control judicial, y que las decisiones basadas exclusivamente en tratamiento automatizado que produzcan efectos jurídicos significativos sobre las personas deben poder ser impugnadas y revisadas por órganos humanos competentes. Esta doctrina es directamente aplicable a las situaciones en que la evidencia algorítmica constituye el fundamento principal de la acusación penal.

El Reglamento de Inteligencia Artificial de la Unión Europea (UE 2024/1689) clasifica como sistemas de alto riesgo aquellos utilizados en el ámbito de la justicia penal para evaluar el riesgo de reincidencia, para asistir en la investigación de delitos o para interpretar evidencia, e impone a sus proveedores requisitos de: documentación técnica completa; transparencia e información suficiente para la supervisión humana; precisión, robustez y ciberseguridad; y sistemas de gestión de riesgos continuos. Aunque el Reglamento no contiene disposiciones procesales penales específicas sobre la admisibilidad de la evidencia generada por estos sistemas, los requisitos de transparencia y documentación que impone constituyen la condición necesaria para que los órganos judiciales puedan evaluar la fiabilidad de dicha evidencia conforme a los estándares del artículo 6 del CEDH (Comisión Europea, 2024; Floridi et al., 2022).

El panorama latinoamericano: regulación incipiente y desafíos urgentes

En el contexto latinoamericano, la regulación de la evidencia digital en el proceso penal se encuentra en una fase de desarrollo significativamente menos avanzada que en Europa o los Estados Unidos, con la excepción de algunas disposiciones de la legislación procesal de Colombia, Chile y Argentina. La ausencia de marcos normativos específicos sobre la admisibilidad de la evidencia algorítmica expone a los sistemas de justicia de la región a los mismos riesgos de erosión de garantías que se observan en los sistemas más avanzados, pero

sin los mecanismos de control que aquellos han comenzado a desarrollar (Palazzi, 2022; Ortega & Salinas, 2023).

En Colombia, el Código de Procedimiento Penal (Ley 906 de 2004) incluye disposiciones sobre la prueba electrónica que han sido objeto de desarrollo jurisprudencial por parte de la Corte Suprema de Justicia, que ha establecido estándares de autenticidad, integridad y pertinencia aplicables a la evidencia digital. Sin embargo, estas disposiciones no contemplan específicamente la evidencia generada por sistemas de IA. En Argentina, la Ley de Delitos Informáticos (Ley 26.388 de 2008) y las reformas procesales recientes han introducido algunas disposiciones sobre la recolección y preservación de evidencia digital, pero igualmente sin alcanzar la especificidad requerida para la evidencia algorítmica. En Ecuador, el Código Orgánico Integral Penal (COIP) establece en sus artículos 498 y siguientes disposiciones generales sobre la prueba documental y pericial que son aplicables, con interpretación extensiva, a la evidencia digital, pero no existe regulación específica sobre la evidencia algorítmica ni sobre los estándares de admisibilidad de las conclusiones de sistemas de IA (Palazzi, 2022; Miranda Estrampes, 2022).

Cadena de Custodia de la Evidencia Digital: Protocolos y Garantías

El concepto de cadena de custodia en el entorno digital

La cadena de custodia (chain of custody) es un concepto procesal fundamental que documenta el historial de control de una evidencia desde su obtención original hasta su presentación en juicio, garantizando que no ha sido alterada, contaminada ni sustituida. En el contexto de la evidencia física, la cadena de custodia se documenta mediante registros de acceso, embalajes sellados y protocolos de transferencia que permiten al tribunal y a las partes verificar la integridad de la evidencia. En el entorno digital, estos mecanismos deben ser adaptados a las características específicas de los datos digitales, especialmente su volatilidad y duplicabilidad (Casey, 2019; Karie & Venter, 2020).

La literatura forense especializada ha establecido un consenso metodológico sobre los elementos indispensables de la cadena de custodia de la evidencia digital: la identificación y documentación exhaustiva de la evidencia en el momento de su obtención (incluyendo la captura de imagen forense del soporte digital mediante herramientas certificadas); la verificación de la integridad mediante funciones hash criptográficas (MD5, SHA-256) que

permiten detectar cualquier modificación posterior; el registro documentado de todos los accesos a la evidencia y de todas las operaciones realizadas sobre ella; el almacenamiento en condiciones de seguridad que prevengan accesos no autorizados y daños físicos; y la trazabilidad completa de la cadena de transferencia de custodia (Reyes & Lutteral, 2021; Karie & Venter, 2020).

Desafíos específicos en la cadena de custodia de la evidencia algorítmica

La evidencia algorítmica presenta desafíos adicionales a los que ya plantea la evidencia digital convencional. En primer lugar, la captura de la evidencia en tiempo real: muchos sistemas de IA operan en tiempo real y generan conclusiones que no son almacenadas de manera persistente salvo que se implementen mecanismos específicos de registro. Si el sistema de reconocimiento facial identifica a un sospechoso en una transmisión de videovigilancia en tiempo real pero no almacena el proceso de análisis que condujo a esa identificación, la posibilidad de revisar y cuestionar esa conclusión a posteriori queda comprometida (Garvie et al., 2022; Maras & Alexandrou, 2023).

En segundo lugar, la documentación del entorno técnico en el momento de la generación de la evidencia: el resultado producido por un sistema de IA puede variar en función de parámetros de configuración, versiones del software, calidad de los datos de entrada y condiciones del entorno de ejecución. Para poder evaluar la fiabilidad de la conclusión algorítmica, es indispensable documentar con exactitud las condiciones técnicas en que fue generada, lo que requiere protocolos de logging y auditoría que no siempre están implementados en los sistemas desplegados operativamente (Raeini et al., 2023; Cheng et al., 2024).

En tercer lugar, la disponibilidad del sistema para su examen pericial: la defensa tiene derecho a que peritos de su parte puedan examinar el sistema que generó la evidencia, replicar las condiciones en que fue generada y someter sus conclusiones a prueba de contraste. Pero en muchos casos los sistemas de IA utilizados por las fuerzas de seguridad son propiedad de empresas privadas que invocan el secreto comercial para impedir el acceso de peritos externos al código fuente y a los datos de entrenamiento del sistema. Esta tensión entre el derecho de defensa y la propiedad intelectual del sistema ha sido objeto de intenso debate

doctrinal y jurisprudencial, especialmente en el contexto estadounidense (Pasquale, 2020; Ferguson, 2021).

Garantías Procesales Penales en Riesgo: Análisis Sistemático

El principio de contradicción y el derecho de defensa

El principio de contradicción, configurado en el derecho continental como pilar del sistema acusatorio y en el common law como el adversarial principle, exige que toda evidencia presentada por la acusación sea puesta a disposición de la defensa para que pueda ser conocida, examinada y refutada. En el contexto de la evidencia algorítmica, este principio se ve amenazado por al menos tres mecanismos. El primero es la invocación del secreto comercial por parte de los proveedores del sistema de IA, que puede impedir a la defensa acceder al código fuente, los datos de entrenamiento y los parámetros de funcionamiento del sistema, elementos sin los cuales resulta prácticamente imposible evaluar la fiabilidad de sus conclusiones (Pasquale, 2020; Završnik, 2021).

El segundo mecanismo de amenaza al principio de contradicción es la complejidad técnica intrínseca de la evidencia algorítmica, que puede hacer que incluso peritos bien cualificados no estén en condiciones de examinar y cuestionar de manera efectiva las conclusiones de sistemas de deep learning de alta complejidad. La asimetría técnica entre los equipos de la fiscalía (que habitualmente tienen acceso privilegiado a los sistemas y a sus proveedores) y los de la defensa (que dependen del acceso que el tribunal les conceda y de los recursos económicos del acusado para contratar peritos especializados) genera una desigualdad de armas (inequality of arms) que el TEDH ha declarado incompatible con el artículo 6 del CEDH cuando alcanza un nivel suficiente de gravedad (Hildebrandt, 2020; Kerr, 2020).

El tercer mecanismo de amenaza es la denominada launderización epistémica de la evidencia algorítmica: el fenómeno por el cual las conclusiones de un sistema de IA, al ser presentadas en juicio revestidas de la aparente objetividad y precisión técnica propias de los sistemas computacionales, pueden adquirir un peso probatorio excesivo que el tribunal no es capaz de contrarrestar con las herramientas normales de valoración crítica de la prueba. Estudios empíricos realizados con jurados simulados en el contexto estadounidense han documentado que las conclusiones de sistemas algorítmicos son percibidas como más fiables y objetivas que

las de los peritos humanos, incluso cuando su tasa de error es comparable o superior (Dressel & Farid, 2022; Cheng et al., 2024).

El principio de presunción de inocencia y el estándar de prueba

La evidencia algorítmica plantea un desafío específico al principio de presunción de inocencia en su dimensión probatoria, que en los sistemas continentales se expresa en la regla de que la carga de la prueba corresponde exclusivamente a la acusación y en que la duda razonable debe beneficiar al acusado. La naturaleza probabilística de muchas conclusiones algorítmicas introduce en el proceso penal una incertidumbre estadística cuya traducción en términos de certeza judicial no ha sido resuelta de manera satisfactoria.

Un sistema de evaluación de riesgo que asigna al acusado un índice de riesgo de reincidencia del 85% no está afirmando que el acusado vaya a reincidir con certeza, sino que, dentro de la población de individuos con el mismo perfil de variables, el 85% reincide. Esta afirmación estadística no puede ser utilizada directamente como prueba de que el acusado en concreto reincidirá, so pena de cometer la denominada falacia del fiscal (prosecutor's fallacy): la confusión entre la probabilidad de que el perfil del acusado corresponda al de un culpable y la probabilidad de que el acusado concreto sea culpable. La jurisprudencia angloamericana ha abordado este problema en el contexto de la prueba de ADN, pero su traslación al ámbito de la evidencia algorítmica es todavía incompleta (Doshi-Velez & Kim, 2021; Dressel & Farid, 2022).

El deber de motivación de las resoluciones judiciales

El deber de motivación de las resoluciones judiciales, reconocido como exigencia constitucional y convencional en la práctica totalidad de los ordenamientos democráticos, impone al juez la obligación de explicar las razones por las que otorga o deniega valor probatorio a cada elemento de prueba. Cuando el elemento de prueba es la conclusión de un sistema de IA opaco, el cumplimiento de este deber de motivación se enfrenta a un obstáculo estructural: el juez no puede explicar por qué el sistema llegó a esa conclusión en el caso concreto si el propio sistema no puede proporcionarle esa explicación.

La literatura especializada ha identificado dos respuestas posibles a este problema. La primera es la exigencia de explicabilidad como condición de admisibilidad: solo los sistemas de IA que puedan proporcionar explicaciones comprensibles para sus conclusiones en el caso concreto

deberían ser admitidos como fuente de evidencia en el proceso penal. Esta posición, defendida con fuerza por Doshi-Velez & Kim (2021) y por Arrieta et al. (2020), tiene la virtud de preservar íntegramente el deber de motivación, pero al precio de excluir del proceso penal algunas de las herramientas algorítmicas más potentes, precisamente aquellas cuya complejidad dificulta la explicabilidad. La segunda respuesta es la exigencia de auditabilidad como sustituto de la explicabilidad: el sistema no necesita ser capaz de explicar cada decisión, pero debe poder ser auditado por peritos independientes que verifiquen su funcionamiento general y su tasa de error en condiciones equivalentes a las del caso concreto. Esta posición, más pragmática, permite el uso de sistemas de alta complejidad pero exige la creación de organismos independientes de auditoría forense algorítmica (Wachter et al., 2020; Floridi et al., 2022).

Modelos de Peritaje Forense Digital en el Proceso Penal

El peritaje forense digital tradicional y sus limitaciones ante la IA

El peritaje forense digital ha evolucionado durante las últimas tres décadas como una disciplina técnico-jurídica especializada, con sus propios estándares metodológicos, herramientas certificadas y marcos de acreditación. Organismos como el Scientific Working Group on Digital Evidence (SWGDE) en Estados Unidos, el Forensic Science International Digital Investigation en el ámbito académico, y los laboratorios forenses acreditados por ENISA en Europa, han establecido marcos metodológicos que abarcan la captura, preservación, análisis y presentación de evidencia digital.

Sin embargo, estos marcos metodológicos fueron desarrollados fundamentalmente en el paradigma de la evidencia digital convencional (dispositivos de almacenamiento, comunicaciones electrónicas, registros de actividad) y presentan limitaciones significativas cuando se enfrentan a la evidencia algorítmica. La verificación de la integridad mediante funciones hash, por ejemplo, garantiza que los datos de entrada al sistema de IA no han sido modificados, pero no dice nada sobre la fiabilidad del proceso mediante el cual el sistema transformó esos datos en sus conclusiones. Los estándares de documentación de la cadena de custodia contemplan los soportes físicos y los ficheros digitales, pero no los modelos de IA, los conjuntos de datos de entrenamiento ni las versiones del software que determinan el comportamiento del sistema (Karie & Venter, 2020; Raeini et al., 2023).

Modelos emergentes: auditoría algorítmica forense

La literatura más reciente ha comenzado a delinear los contornos de un nuevo modelo de peritaje forense especializado en la evaluación de sistemas de IA: la auditoría algorítmica forense. Este modelo emergente propone que la verificación de la fiabilidad de la evidencia algorítmica en el proceso penal se lleve a cabo mediante un protocolo de auditoría que comprenda al menos las siguientes etapas: (1) documentación completa del sistema (arquitectura, datos de entrenamiento, versión del software, parámetros de configuración en el momento de la generación de la evidencia); (2) verificación de la tasa de error general del sistema en condiciones similares al caso concreto; (3) evaluación de los sesgos del sistema respecto de variables relevantes (etnia, género, condición socioeconómica); (4) replicación de la conclusión del sistema en condiciones controladas; y (5) análisis de la explicabilidad local de la conclusión específica mediante herramientas de interpretabilidad como LIME, SHAP u otras técnicas de explicación post-hoc (Arrieta et al., 2020; Cheng et al., 2024; Maras & Alexandrou, 2023).

Este modelo de auditoría algorítmica forense enfrenta, no obstante, obstáculos significativos para su implementación práctica. En primer lugar, exige la disponibilidad del sistema para su examen por auditores independientes, lo que choca con las políticas de propiedad intelectual de los proveedores comerciales. En segundo lugar, requiere la existencia de organismos de auditoría independientes con las capacidades técnicas y los recursos necesarios, que en la actualidad son escasos. En tercer lugar, la rápida evolución de los sistemas de IA hace que los resultados de una auditoría pueden ser insuficientes para evaluar versiones del sistema posteriores a la realización de la auditoría (Raeini et al., 2023; Ferguson, 2021).

Comparativa de marcos normativos sobre evidencia algorítmica en el proceso penal

Jurisdicción	Marco normativo principal	Estándar de admisibilidad	Garantías específicas para evidencia algorítmica	Déficits identificados
Estados Unidos (federal)	Reg. Fed. de Evidencia, R. 702; Daubert (1993); Kumho Tire (1999)	Fiabilidad metodológica; aceptación científica; aplicación válida al caso	Control judicial de gatekeeper; exigencia de validación pericial	Aplicación inconsistente a IA; opacidad de sistemas propietarios; asimetría de recursos entre acusación y defensa
Unión Europea	Art. 6 CEDH; Reglamento UE IA 2024/1689; jurisprudencia TEDH y TJUE	Transparencia; supervisión humana; explicabilidad para sistemas de alto riesgo	Requisitos de documentación técnica; derecho de impugnación; auditoría independiente contemplada	Ausencia de normas procesales penales armonizadas; disparidad entre Estados miembros; implementación pendiente
Reino Unido	Police and Criminal Evidence Act 1984;	Relevancia; fiabilidad; fairness procesal	Escasas disposiciones específicas; aplicación por	Marco anterior a IA; ausencia de estándar específico para

Jurisdicción	Marco normativo principal	Estándar de admisibilidad	Garantías específicas para evidencia algorítmica	Déficits identificados
	Computer Misuse Act 1990; Case law		analogía del derecho probatorio general	evidencia algorítmica; debates post-Brexit sobre convergencia con UE
Colombia	Código de Procedimiento Penal (Ley 906/2004); Ley 527/1999 sobre mensajes de datos	Autenticidad; integridad; pertinencia; legalidad de obtención	Jurisprudencia de la CSJ sobre prueba electrónica; ausencia de regulación específica de IA	Sin regulación específica de evidencia algorítmica; capacidades periciales limitadas; ausencia de organismos de auditoría
Argentina	Código Procesal Penal Federal; Ley 26.388 de delitos informáticos	Idéntico al estándar general de prueba pericial; autenticidad e integridad de soporte digital	Inexistentes en materia de IA; aplicación de normas generales sobre pericia informática	Marco normativo anterior a IA; sin estándares forenses específicos; disparidad entre provincias

Nota: Elaboración propia a partir de Kerr (2020), Cheng et al. (2024), Hildebrandt (2020), Završnik (2021), Palazzi (2022), Miranda Estrampes (2022) y Comisión Europea (2024).

Hallazgos Transversales: Tensiones Estructurales en la Evidencia Algorítmica

La tensión entre eficiencia probatoria y garantías procesales

Un hallazgo transversal de la revisión bibliográfica es la constatación de una tensión estructural entre la eficiencia probatoria que los sistemas de IA pueden proporcionar y las garantías procesales que el Estado de Derecho exige. Los sistemas algorítmicos pueden analizar volúmenes de evidencia digital que serían inmanejables para el análisis humano en tiempos razonables, identificar patrones y conexiones que escaparían a la percepción de los investigadores, y producir conclusiones con niveles de consistencia y reproducibilidad superiores a los del juicio humano en determinadas tareas. Estas capacidades representan una oportunidad genuina para mejorar la calidad epistémica del proceso penal en la determinación de los hechos (Ferguson, 2021; Raeini et al., 2023).

Sin embargo, estas mismas capacidades introducen riesgos específicos para las garantías procesales cuando se emplean sin los controles adecuados. La velocidad y el volumen del análisis algorítmico pueden generar una ilusión de exhaustividad que oculta los sesgos de selección inherentes a los datos de entrenamiento. La aparente objetividad del resultado numérico puede oscurecer la naturaleza fundamentalmente normativa y política de los criterios que el sistema incorpora en sus modelos. Y la complejidad técnica del sistema puede hacer prácticamente imposible la contradicción efectiva por parte de la defensa sin acceso a recursos técnicos y económicos extraordinarios (Angwin et al., 2016; Dressel & Farid, 2022; Završnik, 2021).

El problema del sesgo algorítmico como amenaza a la igualdad procesal

La revisión bibliográfica confirma de manera consistente que los sistemas de IA utilizados en el ámbito de la justicia penal exhiben tasas de error diferenciadas por grupos raciales, de género y socioeconómicos. El estudio fundacional de Angwin et al. (2016) sobre el sistema COMPAS documentó que este sistema clasifica incorrectamente como de alto riesgo a los imputados afroamericanos a una tasa casi doble que la de los imputados blancos, mientras que clasifica incorrectamente como de bajo riesgo a los imputados blancos a una tasa casi doble que la de los imputados afroamericanos. Estudios posteriores de Buolamwini & Gebru

(2020) sobre sistemas de reconocimiento facial documentaron tasas de error de hasta el 34% en la identificación de mujeres de piel oscura, frente al 0.8% en hombres de piel clara. Estos hallazgos, replicados en múltiples contextos geográficos y con distintos sistemas, constituyen evidencia sólida de que el sesgo algorítmico no es un problema marginal o excepcional, sino una característica estructural de los sistemas actuales.

Las consecuencias procesales de este sesgo son de primera magnitud: si un sistema de IA exhibe tasas de error diferenciadas por grupos, su utilización como fundamento de decisiones procesales penales (medidas cautelares, sentencias, decisiones sobre libertad condicional) reproduce y amplifica las desigualdades estructurales del sistema de justicia, confiriéndoles la aparente legitimidad de la objetividad técnica. Este fenómeno ha sido caracterizado como discriminación algorítmica (Završnik, 2021) y constituye una amenaza directa al principio de igualdad ante la ley consagrado en el artículo 14 del CEDH y en los textos constitucionales de la gran mayoría de los ordenamientos democráticos (Comisión Europea, 2024; Floridi et al., 2022).

DISCUSIÓN Y CONCLUSIONES

Discusión de los Hallazgos

Los resultados obtenidos de la revisión bibliográfica y del análisis comparado permiten articular un conjunto de reflexiones críticas que sitúan los hallazgos del presente capítulo en el contexto más amplio de los debates sobre gobernanza de la IA en el ámbito de la justicia penal. La discusión que sigue se organiza en torno a cuatro ejes temáticos: la brecha normativa, la dimensión técnica, la dimensión institucional y la perspectiva latinoamericana.

La brecha normativa y sus consecuencias

El hallazgo más robusto y consistente del presente estudio es la constatación de una brecha normativa de primera magnitud entre la sofisticación técnica de los sistemas generadores de evidencia algorítmica y la capacidad del derecho procesal penal para someterlos a los controles de legitimidad que el Estado de Derecho exige. Esta brecha no es meramente cuantitativa (el derecho va rezagado respecto a la tecnología) sino cualitativa: las categorías conceptuales del derecho probatorio fueron diseñadas para tipos de evidencia (física, testimonial, documental, pericial) cuyas características epistémicas son fundamentalmente distintas de las de la evidencia algorítmica.

Esta constatación es consistente con los hallazgos de Hildebrandt (2020), quien ha acuñado el concepto de «código como derecho» (code as law) para describir el fenómeno por el cual las decisiones de diseño de los sistemas tecnológicos adquieren una función normativa de facto que escapa a los controles democráticos ordinarios. En el contexto de la evidencia algorítmica, este fenómeno se manifiesta en el hecho de que son los desarrolladores de los sistemas los que, al diseñar los criterios de clasificación y predicción, determinan qué variables son relevantes para la evaluación del riesgo delictivo o para la identificación del sospechoso; y estas decisiones tienen consecuencias directas sobre los derechos fundamentales de los ciudadanos sin haber sido sometidas al escrutinio democrático ni judicial que correspondería (Pasquale, 2020; Završnik, 2021).

El contraste con los trabajos previos de Kerr (2020) y Casey (2019) resulta significativo: mientras que estos autores describían el problema fundamentalmente en términos de inadecuación técnica de las normas existentes (reglas de autenticación, estándares de fiabilidad pericial) que podrían ser adaptadas mediante interpretación o reforma incremental, los estudios más recientes de Cheng et al. (2024) y Raeini et al. (2023) apuntan a una inadecuación más profunda que requiere la construcción de nuevas categorías conceptuales, no meramente la adaptación de las existentes. El presente estudio se alinea con esta posición más ambiciosa, al constatar que los principios de contradicción, motivación y tutela judicial efectiva no pueden ser garantizados mediante la simple extensión de los marcos probatorios tradicionales a la evidencia algorítmica.

La dimensión técnica: entre la explicabilidad y la auditabilidad

Un segundo eje de discusión relevante se refiere a la tensión técnica entre la explicabilidad y la auditabilidad como condiciones alternativas para la admisibilidad de la evidencia algorítmica. La posición de Doshi-Velez & Kim (2021) y Arrieta et al. (2020), que defienden la explicabilidad como requisito indispensable, tiene la virtud de preservar íntegramente el principio de motivación judicial y el derecho de defensa a cuestionar efectivamente la evidencia. Sin embargo, tiene también el costo de excluir del proceso penal los sistemas de IA más potentes y, paradójicamente, potencialmente más precisos, precisamente porque su complejidad dificulta la explicabilidad.

La alternativa de la auditabilidad, defendida por Wachter et al. (2020) y parcialmente adoptada por el Reglamento de IA de la UE, ofrece un equilibrio más pragmático entre las exigencias técnicas y las garantías procesales. Sin embargo, plantea el problema de la independencia y la capacidad técnica de los organismos de auditoría: si los auditores son contratados por las empresas proveedoras del sistema, su independencia resulta cuestionable; si son órganos públicos, su capacidad técnica puede ser insuficiente para evaluar sistemas en rápida evolución. Las propuestas más recientes de Cheng et al. (2024) sobre auditoría algorítmica adversarial, que implica la creación de equipos de auditores tanto de la acusación como de la defensa que compiten en el proceso de evaluación del sistema, constituyen una alternativa interesante que merece mayor desarrollo en investigaciones futuras.

La dimensión institucional: competencias y formación

Un hallazgo que la discusión de los resultados debe subrayar es la brecha institucional entre la complejidad técnica de la evidencia algorítmica y la preparación de los operadores jurídicos (jueces, fiscales, abogados defensores) para evaluarla críticamente. Los estudios empíricos citados en la sección de resultados (Dressel & Farid, 2022; Cheng et al., 2024) documentan que incluso los operadores jurídicos con formación universitaria y experiencia profesional significativa tienen dificultades para comprender y cuestionar las conclusiones de los sistemas de IA, lo que genera el fenómeno de la launderización epistémica descrito anteriormente. Esta brecha institucional no puede ser resuelta exclusivamente mediante reformas normativas, sino que exige programas de formación continua específica para los operadores jurídicos, la creación de unidades de análisis técnico en fiscalías y defensorías públicas, y la incorporación de peritos forenses digitales especializados en IA como parte de la estructura ordinaria del proceso penal.

La perspectiva latinoamericana: urgencia y oportunidad

La situación de los países latinoamericanos respecto a la evidencia algorítmica en el proceso penal presenta una combinación paradójica de urgencia y oportunidad. La urgencia deriva del hecho de que el uso de evidencia digital de todo tipo (incluyendo evidencia generada o procesada algorítmicamente) se ha convertido en una práctica corriente en los sistemas de justicia penal de la región, sin que existan marcos normativos específicos que garanticen su

admisibilidad con las debidas garantías. La oportunidad reside en la posibilidad de aprender de los errores y las insuficiencias de los sistemas más avanzados, diseñando desde el principio marcos normativos que incorporen los estándares de garantía más exigentes sin necesidad de pasar por el ciclo de reforma reactiva que ha caracterizado la respuesta de los sistemas angloamericano y europeo.

En este sentido, los resultados del presente estudio confirman y amplían las observaciones de Ortega & Salinas (2023) sobre la importancia de que los países latinoamericanos no limiten su política regulatoria de la IA a la adopción mimética de los marcos europeos o estadounidenses, sino que desarrollen respuestas propias que tengan en cuenta las especificidades de sus sistemas de justicia, las características de sus problemas de inseguridad, y las particularidades de sus poblaciones y sus tradiciones jurídicas. En el ámbito procesal penal, esto implica reformas que fortalezcan las capacidades periciales públicas, que establezcan estándares de admisibilidad de la evidencia algorítmica adaptados al sistema acusatorio oral prevalente en la región, y que garanticen la igualdad de armas entre acusación y defensa en el acceso a los recursos técnicos necesarios para evaluar la evidencia algorítmica.

Conclusiones

De la revisión y el análisis efectuados en el presente capítulo es posible extraer las siguientes conclusiones, formuladas con la cautela epistemológica que impone la naturaleza documental de la investigación y el estado todavía abierto de muchos de los debates examinados.

La evidencia algorítmica constituye una categoría probatoria cualitativamente distinta de la prueba digital convencional, que requiere un estatuto procesal específico. Sus características diferenciales (opacidad interna, naturaleza probabilística de sus conclusiones, dependencia de datos de entrenamiento potencialmente sesgados, y necesidad de auditoría especializada para la evaluación de su fiabilidad) la hacen incompatible con la simple extensión de los marcos probatorios diseñados para la prueba física, testimonial o documental. Los sistemas jurídicos que carezcan de este estatuto específico están expuestos a un riesgo sistemático de erosión de las garantías procesales penales fundamentales.

La brecha normativa entre la sofisticación técnica de los sistemas generadores de evidencia y la capacidad del derecho procesal penal para someterlos a control es una característica estructural del momento presente, que ningún ordenamiento jurídico ha logrado superar de

manera satisfactoria. El Reglamento de IA de la Unión Europea de 2024 constituye el avance más significativo en este sentido, al imponer requisitos de transparencia, documentación y supervisión humana a los sistemas de IA de alto riesgo utilizados en el ámbito de la justicia penal, pero no contiene disposiciones procesales penales específicas sobre la admisibilidad y valoración de la evidencia generada por dichos sistemas.

El principio de contradicción, el principio de presunción de inocencia y el deber de motivación judicial son las garantías procesales penales más directamente amenazadas por el uso sin controles adecuados de la evidencia algorítmica. Su protección efectiva requiere: que el acusado tenga acceso real a la información técnica necesaria para cuestionar la fiabilidad del sistema; que el sistema pueda ser auditado por peritos independientes en el marco del proceso; que el tribunal sea capaz de motivar su valoración de la evidencia algorítmica con referencia a criterios verificables de fiabilidad; y que las conclusiones del sistema no sean utilizadas de manera determinante sin supervisión humana crítica.

El sesgo algorítmico documentado en los sistemas de evaluación de riesgo y reconocimiento facial utilizados en el ámbito de la justicia penal constituye una amenaza sistémica al principio de igualdad ante la ley que ningún ordenamiento ha logrado neutralizar de manera satisfactoria. Su abordaje efectivo requiere la combinación de medidas técnicas (metodologías de detección y corrección de sesgos), normativas (prohibición de uso de sistemas que exhiban tasas de error discriminatoriamente diferenciadas sin proceso de corrección verificado) e institucionales (mecanismos de impugnación de las decisiones basadas en sistemas sesgados).

El modelo de auditoría algorítmica forense, todavía en proceso de consolidación, ofrece la respuesta más prometedora al problema de la verificación de la fiabilidad de la evidencia algorítmica en el proceso penal. Su implementación práctica requiere la creación de organismos independientes de auditoría con las capacidades técnicas necesarias, la adopción de protocolos estandarizados de auditoría forense de sistemas de IA, y el desarrollo de marcos regulatorios que impongan a los proveedores de sistemas la obligación de facilitar el acceso para fines de auditoría forense.

Los países latinoamericanos se encuentran ante una oportunidad singular para diseñar marcos normativos sobre evidencia algorítmica que incorporen desde el principio los estándares de garantía más exigentes, superando la lógica de reforma reactiva que ha caracterizado la

respuesta de los sistemas más avanzados. Este esfuerzo regulatorio debe articularse en torno a los principios de igualdad de armas, supervisión humana efectiva, explicabilidad proporcional al impacto de la decisión y acceso a recursos técnicos especializados para la defensa.

Entre las limitaciones del presente estudio que deben orientar la investigación futura, destaca la necesidad de análisis empíricos sobre el impacto real del uso de la evidencia algorítmica en las decisiones de los tribunales penales latinoamericanos, la comparación sistemática de las distintas propuestas de reforma legislativa en curso, y el desarrollo de estudios cuantitativos sobre la distribución de las tasas de error de los sistemas utilizados en la región por grupos sociodemográficos. La intersección entre prueba digital y garantías constitucionales en el proceso penal constituye un campo de investigación en el que la producción académica sigue siendo insuficiente respecto a la urgencia de las preguntas planteadas.

CAPITULO V

Ética, Derechos Humanos y Justicia Penal Automatizada

Ética, Derechos Humanos y Justicia Penal Automatizada

TANIA EVELIN MEJIA AUZ[5]

RESUMEN

El presente capítulo analiza la intersección entre la ética aplicada, los derechos humanos internacionalmente reconocidos y los procesos de automatización de la justicia penal mediante sistemas de inteligencia artificial. Partiendo de los fundamentos conceptuales, los problemas de responsabilidad, las tipologías delictivas y la evidencia algorítmica desarrollados en los capítulos precedentes de esta obra, se examinan los marcos éticos y de derechos humanos aplicables a los sistemas automatizados de toma de decisiones penales, los principios de diseño ético de la IA en la justicia criminal, las obligaciones de los Estados en materia de supervisión humana y rendición de cuentas, y los modelos emergentes de gobernanza ético-jurídica de la IA penal. A partir de una revisión bibliográfica sistemática y del análisis comparado de marcos normativos internacionales, regionales y nacionales, se constata que la automatización de la justicia penal plantea riesgos específicos para los derechos a la no discriminación, la dignidad humana, el debido proceso, la presunción de inocencia y la tutela judicial efectiva. Se concluye que la compatibilidad de la justicia penal automatizada con el Estado de Derecho exige marcos éticos vinculantes, mecanismos de rendición de cuentas algorítmica y garantías de supervisión humana significativa que la normativa vigente no garantiza plenamente.

Palabras clave: *ética de la inteligencia artificial; derechos humanos; justicia penal automatizada; supervisión humana; discriminación algorítmica; rendición de cuentas; debido proceso; gobernanza algorítmica.*

[5] Abogada de los Juzgados y Tribunales de la República del Ecuador, Universidad Guayaquil, Especilista en Ciencias Penales y Criminologicas, Universidad de Guayaquil, Magister en Derecho Constitucional, Universidad Tecnologica Empresarial de Guayaquil, Magister en Derecho Mención en Derecho Procesal Penal, Universidad Estatal de Milagro, Funcionaria Publica, Fiscalía General del Estado, correo electrónico: ab.taniamejiauz@hotmail.com.

ABSTRACT

This chapter analyzes the intersection between applied ethics, internationally recognized human rights, and the automation of criminal justice through artificial intelligence systems. Building on the conceptual foundations, liability issues, criminal typologies, and algorithmic evidence developed in the preceding chapters of this work, it examines the ethical and human rights frameworks applicable to automated criminal decision-making systems, the principles of ethical AI design in criminal justice, States' obligations regarding human oversight and accountability, and emerging models of ethical-legal governance of criminal AI. Based on a systematic literature review and comparative analysis of international, regional, and national normative frameworks, the chapter finds that the automation of criminal justice poses specific risks to the rights to non-discrimination, human dignity, due process, presumption of innocence, and effective judicial protection. It concludes that the compatibility of automated criminal justice with the Rule of Law requires binding ethical frameworks, algorithmic accountability mechanisms, and meaningful human oversight guarantees that existing regulations do not yet fully secure.

Keywords: *artificial intelligence ethics; human rights; automated criminal justice; human oversight; algorithmic discrimination; accountability; due process; algorithmic governance.*

INTRODUCCIÓN

Los capítulos precedentes de esta obra han trazado un panorama sistemático de los desafíos que la inteligencia artificial plantea al derecho penal contemporáneo: desde los fundamentos conceptuales que articulan la intersección entre tecnología y ius puniendi (Capítulo I), pasando por la crisis de imputación que los sistemas autónomos generan en las categorías clásicas de la responsabilidad penal (Capítulo II) y las nuevas tipologías delictivas que la IA hace posibles o transforma cualitativamente (Capítulo III), hasta la problemática de la evidencia algorítmica en el juicio penal (Capítulo IV). El presente capítulo avanza en la secuencia lógica de la obra abordando la dimensión que subyace y trasciende a todas las anteriores: la fundamentación ética y de derechos humanos de la justicia penal automatizada, y los marcos normativos emergentes que intentan dar respuesta a esa fundamentación.

La necesidad de este análisis no es académica ni programática: es urgente y concreta. En el momento en que se redactan estas páginas, miles de decisiones penales relevantes se adoptan diariamente en distintos países con apoyo determinante de sistemas algorítmicos: decisiones sobre prisión preventiva que afectan a la libertad física de las personas, decisiones sobre libertad condicional que determinan si un ser humano permanece o no encarcelado, decisiones sobre vigilancia policial que concentran los recursos del Estado sobre determinadas poblaciones y territorios, decisiones sobre deportación con consecuencias irreversibles para la vida de personas y familias. La pregunta de si estas decisiones son éticamente legítimas y compatibles con los derechos humanos es, por tanto, una pregunta con consecuencias reales e inmediatas sobre vidas humanas concretas (Angwin et al., 2016; Završnik, 2021; Floridi et al., 2022).

La relevancia teórica del problema se asienta sobre una paradoja que la literatura especializada ha señalado con creciente énfasis: los sistemas algorítmicos son introducidos en la justicia penal precisamente con la promesa de reducir la arbitrariedad y los sesgos subjetivos propios del juicio humano, de incrementar la consistencia y la objetividad de las decisiones, y de mejorar la eficiencia de un sistema sobrecargado. Sin embargo, la evidencia empírica acumulada durante la última década muestra que, en la práctica, estos sistemas con frecuencia reproducen, amplifican y otorgan una apariencia de legitimidad técnica a las desigualdades estructurales y los sesgos raciales, de género y de clase que caracterizan históricamente a los sistemas de justicia penal (Angwin et al., 2016; Buolamwini & Gebru,

2020; Dressel & Farid, 2022). Esta paradoja sitúa la pregunta ética en el centro del debate: ¿puede la automatización de la justicia penal ser compatible con los principios de igualdad, dignidad humana y no discriminación que los sistemas democráticos proclaman como fundamentos del Estado de Derecho?

La investigación en ética de la inteligencia artificial ha experimentado un desarrollo extraordinario durante la última década, generando una pluralidad de marcos teóricos y de principios orientadores que compiten por definir los estándares de diseño y despliegue éticamente responsable de los sistemas de IA. Los trabajos fundacionales de Floridi et al. (2018) sobre los principios éticos de la IA, de Jobin et al. (2019) sobre la convergencia y divergencia de los marcos éticos globales, de Mittelstadt et al. (2016) sobre la ética de los algoritmos, y de Zuboff (2019) sobre el capitalismo de vigilancia han configurado el horizonte conceptual del debate. En el ámbito específico de la justicia penal, los trabajos de Završnik (2021) sobre la justicia penal algorítmica en Europa, de Ferguson (2021) sobre el policing predictivo, de Richardson et al. (2019) sobre los datos contaminados en el sistema de justicia, y de Pasquale (2015; 2020) sobre la sociedad de la caja negra han aportado los análisis más rigurosos y pertinentes. Desde la perspectiva del derecho internacional de los derechos humanos, los informes del Relator Especial de las Naciones Unidas sobre promoción y protección del derecho a la privacidad (Cannataci, 2021), de la Comisión Interamericana de Derechos Humanos (CIDH, 2021) sobre tecnología y derechos humanos, y del Consejo de Derechos Humanos (CDH, 2023) sobre el impacto de la IA en los derechos humanos, constituyen los referentes normativos más autorizados en el ámbito internacional.

El propósito del presente capítulo es, en consecuencia, cuádruple. En primer lugar, examinar los marcos éticos internacionales aplicables a la inteligencia artificial en la justicia penal, evaluando su coherencia y su capacidad de orientación normativa. En segundo lugar, analizar los derechos humanos específicamente afectados por la automatización de la justicia penal, tanto en la fase de investigación como en la de enjuiciamiento y ejecución de penas. En tercer lugar, identificar los principios de diseño ético de sistemas de IA para la justicia penal que emergen de la convergencia entre la ética aplicada y el derecho internacional de los derechos humanos. En cuarto lugar, evaluar los modelos de gobernanza ético-jurídica de la IA penal que distintos ordenamientos han comenzado a articular, identificando sus fortalezas, debilidades y las reformas necesarias para garantizar la compatibilidad de la justicia penal automatizada

con el Estado de Derecho. Este análisis no reitera los desarrollos efectuados en los capítulos precedentes sobre responsabilidad, tipología delictiva y evidencia, sino que los presupone y los proyecta sobre el plano específico de la fundamentación ética y de derechos humanos, configurando el marco normativo de referencia para el capítulo conclusivo de la obra.

MATERIAL Y MÉTODOS

Diseño de investigación

El presente capítulo adopta un diseño de investigación de naturaleza teórico-documental, coherente con la estrategia metodológica general de la obra de la que forma parte y con los diseños adoptados en los capítulos precedentes. El diseño combina la revisión bibliográfica sistemática con el método comparativo de fuentes secundarias, abordando la literatura científica especializada en ética de la inteligencia artificial, derecho internacional de los derechos humanos, derecho penal comparado y gobernanza tecnológica; los instrumentos normativos internacionales, regionales y nacionales aplicables; los documentos de política pública de organismos internacionales relevantes; y la jurisprudencia comparada pertinente en materia de derechos fundamentales y tecnología. Esta combinación metodológica resulta la más adecuada para el objeto de estudio, dado que los problemas éticos y de derechos humanos que plantea la justicia penal automatizada son eminentemente conceptuales, normativos y comparativos, y su análisis riguroso no requiere ni se beneficiaría de un trabajo de campo cuantitativo en el estado actual de desarrollo del campo (Siems, 2018; Zweigert & Kötz, 2020).

El método comparativo se emplea en su modalidad funcional, partiendo de los problemas éticos y de derechos humanos que los sistemas automatizados de justicia penal plantean de manera común en distintos ordenamientos y contextos culturales, para examinar después las respuestas que cada sistema jurídico y cada marco de gobernanza ofrece a esos problemas. Esta aproximación permite identificar convergencias y divergencias significativas entre los modelos normativos y extraer principios de diseño ético aplicables a contextos diversos, incluyendo el latinoamericano, que ocupa un lugar de atención específica en la obra.

La metodología se enmarca en la tradición investigadora del derecho comparado y de la filosofía del derecho, con apertura transdisciplinaria a los estudios de tecnología y sociedad, la ética aplicada y la ciencia política. Esta apertura es necesaria dado que los problemas

analizados no pueden ser abordados satisfactoriamente desde ninguna disciplina aislada: la comprensión de los sesgos algorítmicos requiere conocimiento de las ciencias de la computación; la evaluación de su impacto sobre los derechos humanos requiere jurisprudencia internacional; la propuesta de marcos de gobernanza requiere ciencia política; y la fundamentación de los estándares éticos requiere filosofía moral. La investigación en este campo es, necesariamente, interdisciplinaria.

Fuentes y corpus documental

La construcción del corpus documental se realizó a través de búsquedas sistemáticas en las bases de datos Scopus, Web of Science, HeinOnline, PhilPapers y SSRN durante el período comprendido entre febrero de 2024 y marzo de 2025. Los descriptores de búsqueda incluyeron los términos: "AI ethics criminal justice", "human rights artificial intelligence", "algorithmic accountability criminal law", "automated decision-making due process", "algorithmic discrimination criminal", "AI governance human rights", "predictive policing human rights", "fairness AI criminal justice", y sus equivalentes en español, francés y alemán. Se priorizaron artículos de revistas indexadas en cuartiles Q1 y Q2 según el Journal Citation Reports, así como monografías de editoriales académicas de reconocido prestigio en los ámbitos de la filosofía del derecho, la ética de la tecnología y el derecho internacional de los derechos humanos.

Se incorporaron también al corpus los siguientes instrumentos normativos e institucionales de primer orden: la Recomendación de la UNESCO sobre la Ética de la Inteligencia Artificial (2021); los Principios de la OCDE sobre Inteligencia Artificial (2023); las Directrices del Consejo de Europa sobre inteligencia artificial en el sistema judicial (CM/Rec(2020)1); el Marco de Ética de la IA de la Unión Africana (2024); el Reglamento de Inteligencia Artificial de la Unión Europea (2024/1689); los informes del Relator Especial de Naciones Unidas sobre privacidad (2021, 2023); los informes del Alto Comisionado de las Naciones Unidas para los Derechos Humanos sobre IA y derechos humanos (2021, 2023); la Resolución 47/22 del Consejo de Derechos Humanos sobre nuevas tecnologías y derechos humanos (2021); el Informe temático de la Comisión Interamericana de Derechos Humanos sobre tecnología y derechos humanos (CIDH, 2021); y los marcos de regulación de IA de Brasil, Colombia, Chile, Argentina, Ecuador y México, en su estado de avance a la fecha de cierre del corpus. El corpus final procesado comprende 189 fuentes documentales, de las cuales 48 fueron seleccionadas como

referencias directamente citadas en el texto, atendiendo a criterios de pertinencia temática, rigor metodológico, representatividad geográfica y actualidad.

Procesamiento y análisis

El análisis de los materiales seleccionados se llevó a cabo mediante la técnica del análisis documental cualitativo, complementada con la construcción de categorías analíticas propias a partir de la revisión de la literatura. La organización de los marcos éticos y de derechos humanos aplicables a la justicia penal automatizada se elaboró de manera inductivo-deductiva: inductiva en la identificación de los problemas concretos que la literatura empírica y doctrinal señala; deductiva en la evaluación de esos problemas a la luz de los principios éticos y de derechos humanos que los instrumentos internacionales establecen. Esta sistematización se comparó con los marcos analíticos propuestos por los autores de referencia (Floridi et al., 2022; Jobin et al., 2019; Mittelstadt et al., 2016) para verificar su consistencia y completitud.

Entre las principales limitaciones del diseño metodológico conviene señalar las siguientes. La acelerada evolución del campo normativo —con nuevas regulaciones de IA en proceso de adopción en múltiples jurisdicciones en el momento de cierre del corpus— implica que algunos aspectos de la comparación normativa pueden haber sido superados por desarrollos posteriores. La preponderancia de la literatura académica en inglés y en el contexto angloamericano y europeo occidental introduce un sesgo de selección que se ha intentado compensar con la incorporación sistemática de fuentes doctrinales y documentos institucionales de América Latina, África y Asia. Por último, la naturaleza normativa del análisis ético implica que las propuestas formuladas tienen un carácter orientativo y de lege ferenda, no de descripción positiva del derecho vigente.

RESULTADOS

Marcos éticos internacionales de la inteligencia artificial y su aplicación a la justicia penal

La convergencia y divergencia de los marcos éticos globales

La revisión de la literatura sobre ética de la inteligencia artificial revela un panorama de notable proliferación de marcos, principios y directrices que, a pesar de sus diferencias de énfasis y de fundamento filosófico, convergen en torno a un núcleo de valores compartidos. El análisis sistemático de Jobin et al. (2019) de 84 documentos de principios éticos de IA

publicados entre 2016 y 2019 identificó cinco principios que aparecen en más del 70% de los documentos analizados: transparencia, justicia y equidad, no maleficencia, responsabilidad (accountability) y privacidad. Estudios más recientes de Hagendorff (2020) y de Canca (2020) han confirmado y enriquecido esta convergencia, añadiendo principios de dignidad humana, sostenibilidad, solidaridad y autonomía que emergen con mayor frecuencia en los marcos elaborados fuera del ámbito angloamericano.

Sin embargo, la convergencia nominal en torno a estos principios oculta divergencias sustantivas en su interpretación, jerarquización y mecanismos de implementación. La interpretación del principio de transparencia varía desde la exigencia de explicabilidad técnica completa del funcionamiento del sistema (posición angloamericana dominante) hasta la exigencia de transparencia procedimental sobre los procesos de diseño y validación (posición de algunos marcos europeos y latinoamericanos). La interpretación del principio de equidad enfrenta tensiones irresueltas entre la equidad estadística (que exige paridad matemática en las métricas de error entre grupos) y la equidad individual (que exige que cada persona sea juzgada conforme a sus propias características y no en función de estadísticas grupales), tensión que Chouldechova (2017) ha demostrado que es matemáticamente irresoluble en determinadas condiciones. Y la interpretación del principio de responsabilidad varía desde la responsabilidad moral subjetiva (que busca identificar al autor moralmente responsable) hasta la responsabilidad sistémica (que distribuye obligaciones de rendición de cuentas entre todos los actores del ecosistema de la IA) (Mittelstadt et al., 2016; Floridi et al., 2022).

En el ámbito específico de la justicia penal, el marco ético más influyente a nivel global es la Recomendación de la UNESCO sobre la Ética de la Inteligencia Artificial, adoptada por aclamación por los 193 Estados miembros en noviembre de 2021. Este instrumento, el primero de alcance verdaderamente global en materia de ética de la IA, consagra como principios centrales la proporcionalidad, la seguridad, la equidad y la no discriminación, la sostenibilidad, el derecho a la privacidad y la protección de datos, la supervisión y control humanos, la transparencia y la explicabilidad, la responsabilidad (accountability), y la protección de los derechos humanos. En su aplicación al ámbito de la justicia, la Recomendación establece que los sistemas de IA que afecten a los derechos individuales deben estar sujetos a supervisión humana significativa, que las decisiones con efectos jurídicos relevantes no pueden adoptarse exclusivamente sobre la base de sistemas

algorítmicos, y que los Estados deben garantizar el derecho de las personas afectadas a conocer que han sido objeto de una decisión algorítmica y a impugnarla de manera efectiva (UNESCO, 2021).

El Marco de los Principios de la OCDE sobre Inteligencia Artificial, adoptado en 2019 y revisado en 2023 para incorporar los sistemas de IA generativa, establece cinco principios rectores: crecimiento inclusivo, desarrollo sostenible y bienestar; valores centrados en el ser humano y equidad; transparencia y explicabilidad; robustez, seguridad y protección; y rendición de cuentas. En su aplicación al ámbito de la justicia penal, la OCDE (2023) ha subrayado la necesidad de que los sistemas de IA utilizados en decisiones de alto impacto sobre las personas sean auditables, que sus resultados sean revisables por órganos humanos competentes, y que existan mecanismos de recurso efectivos para las personas afectadas. El Consejo de Europa, a través de su Recomendación CM/Rec(2020)1 sobre el uso de la inteligencia artificial en los sistemas judiciales, ha añadido el principio de independencia judicial: la IA puede asistir al juez pero no puede sustituir su valoración independiente ni interferir con el ejercicio de la jurisdicción.

Un desarrollo normativo de singular importancia para el presente análisis es la adopción, en el ámbito de la Unión Europea, del Reglamento de IA (2024/1689), que constituye el primer instrumento jurídicamente vinculante de carácter general sobre ética y regulación de la IA a nivel regional. Este Reglamento, que entrará en plena aplicación de manera progresiva entre 2025 y 2027, clasifica como de riesgo inaceptable y por tanto, prohibidos, los sistemas de IA que utilicen técnicas subliminales o manipuladoras para distorsionar el comportamiento de las personas, los sistemas de puntuación social aplicados por autoridades públicas, y ciertos usos de identificación biométrica remota en tiempo real en espacios públicos con fines de represión penal. Como sistemas de alto riesgo, que pueden desplegarse pero solo cumpliendo estrictos requisitos, clasifica los sistemas utilizados para la evaluación de riesgo de reincidencia, los sistemas de identificación biométrica, los sistemas de análisis de evidencia forense y los sistemas de asistencia a la investigación criminal (Comisión Europea, 2024).

Tensiones filosóficas fundamentales en la ética de la IA penal

Más allá de la convergencia nominal en torno a principios compartidos, la revisión de la literatura identifica un conjunto de tensiones filosóficas fundamentales que estructuran el

debate sobre la ética de la IA en la justicia penal y que ningún marco normativo ha resuelto de manera plenamente satisfactoria. Estas tensiones no son meramente académicas: determinan directamente las opciones de diseño, despliegue y regulación de los sistemas algorítmicos en el ámbito penal.

La primera tensión fundamental es la que opone el enfoque consecuencialista al enfoque deontológico. El consecuencialismo evalúa los sistemas algorítmicos en función de sus resultados: si un sistema de predicción de reincidencia reduce efectivamente la comisión de delitos violentos, aunque genere tasas de error diferenciadas por grupos, el consecuencialismo utilitario podría justificarlo por la maximización del bienestar agregado. El deontologismo, en cambio, insiste en que determinadas conductas son intrínsecamente incorrectas con independencia de sus consecuencias: utilizar a una persona como medio para la seguridad colectiva, privarla de libertad en función de estadísticas grupales que no reflejan su culpabilidad individual, viola la dignidad humana con independencia de los beneficios sociales que pudiera generar (Kant, 1785; Rawls, 1971; Floridi et al., 2022). En la práctica de la justicia penal automatizada, esta tensión se manifiesta en el debate sobre los sistemas actuariales de evaluación de riesgo: ¿puede justificarse su uso en decisiones de prisión preventiva porque estadísticamente mejoran la predicción de reincidencia, aunque esto implique tratar a individuos como representantes estadísticos de sus grupos demográficos?

La segunda tensión fundamental opone el principio de autonomía individual al principio de seguridad colectiva. La justicia penal liberal ha construido sus garantías fundamentales sobre el principio de que cada individuo debe ser juzgado por sus propias acciones y no por las acciones o características de su grupo de referencia. La lógica actuarial que subyace a los sistemas de evaluación de riesgo algorítmico invierte parcialmente esta lógica: predice el comportamiento futuro del individuo en función de las características estadísticas del grupo al que pertenece. Esta inversión lógica entra en conflicto no solo con el principio de culpabilidad, sino con la propia concepción del ser humano como agente moral autónomo capaz de determinarse a sí mismo más allá de sus condicionamientos estadísticos (Rawls, 1971; Nussbaum, 2006; Dressel & Farid, 2022).

La tercera tensión fundamental se produce entre la promesa de objetividad del algoritmo y el riesgo de su arbitrariedad encubierta. Uno de los argumentos principales en favor del uso de sistemas algorítmicos en la justicia penal es que eliminan o reducen la subjetividad y los sesgos

cognitivos del juzgador humano (anchoring bias, confirmation bias, implicit racial bias). Sin embargo, la literatura empírica ha demostrado de manera consistente que los sistemas algorítmicos no eliminan los sesgos, sino que los codifican y automatizan, confiriéndoles una apariencia de objetividad técnica que hace más difícil su identificación e impugnación que los sesgos humanos más visibles (Eubanks, 2018; O'Neil, 2016; Richardson et al., 2019). Esta paradoja convierte la promesa de objetividad del algoritmo en uno de sus riesgos más graves: no la transparencia del sesgo, sino su ocultamiento bajo el manto de la neutralidad matemática.

Principios éticos en marcos internacionales de IA y su aplicación a la justicia penal

Principio ético	UNESCO (2021)	OCDE (2023)	Consejo de Europa CM/Rec(2020)1	Reglamento UE IA (2024/1689)
Dignidad humana y no discriminación	Art. 4.1; prohíbe usos que violen DDHH	Principio 1.2; equidad y no discriminación	3.3.1; no discriminación algorítmica en tribunales	Art. 5.1; prohíbe sistemas de puntuación social y manipulación subliminal
Transparencia y explicabilidad	Art. 4.7; derecho a información sobre decisiones algorítmicas	Principio 1.3; obligación de explicabilidad	3.2.2; explicabilidad de las conclusiones judiciales algorítmicas	Arts. 13-14; requisitos de transparencia para sistemas de alto riesgo
Supervisión humana significativa	Art. 4.6; control humano efectivo de	Principio 1.5; rendición de cuentas y supervisión	3.1; el juez no puede ser sustituido por IA	Art. 14; supervisión humana obligatoria en

Principio ético	UNESCO (2021)	OCDE (2023)	Consejo de Europa CM/Rec(2020)1	Reglamento UE IA (2024/1689)
	sistemas de alto impacto		en decisiones jurisdiccionales	sistemas de alto riesgo
Responsabilidad (accountability)	Art. 4.8; responsabilidad de desarrolladores y desplegadores	Principio 1.5; responsabilidad diferenciada en la cadena de IA	3.4; mecanismos de recurso y revisión judicial de decisiones algorítmicas	Arts. 16-29; obligaciones de los proveedores y desplegadores de sistemas de alto riesgo
Privacidad y protección de datos	Art. 4.5; privacidad como requisito de diseño	Principio 1.2; protección de datos y privacidad	3.2.3; proporcionalidad en el uso de datos personales en procesos penales	Arts. 9-11; prohibición de datos sensibles en sistemas de alto riesgo salvo excepciones

Nota: Elaboración propia a partir de UNESCO (2021); OCDE (2023); Consejo de Europa (2020); Comisión Europea (2024).

Derechos humanos específicamente afectados por la justicia penal automatizada

El derecho a la no discriminación y la igualdad ante la ley

El derecho a la no discriminación, consagrado en el artículo 2 de la Declaración Universal de Derechos Humanos, en el artículo 14 del Pacto Internacional de Derechos Civiles y Políticos (PIDCP), en el artículo 14 del Convenio Europeo de Derechos Humanos y en los textos constitucionales de la práctica totalidad de los Estados democráticos, es el derecho más directa y sistemáticamente amenazado por la automatización de la justicia penal. La amenaza no es hipotética: está documentada empíricamente con una solidez que la hace prácticamente irrebatible.

El estudio fundacional de Angwin et al. (2016), publicado en ProPublica bajo el título "Machine Bias", documentó que el sistema COMPAS, utilizado en más de dos tercios de los estados de la Unión Americana para fundamentar decisiones judiciales de prisión preventiva, libertad condicional y sentencia, clasifica incorrectamente como de alto riesgo a los imputados afroamericanos a una tasa casi el doble que la de los imputados blancos, y clasifica incorrectamente como de bajo riesgo a los imputados blancos a una tasa casi el doble que la de los afroamericanos. Este hallazgo fue confirmado con metodologías independientes por Dressel y Farid (2022), quienes además demostraron que la capacidad predictiva del sistema no es significativamente superior a la de personas sin formación específica cuando se les proporciona información básica sobre el caso. La conclusión, que el sistema combina menor precisión con mayor sesgo racial, es de una gravedad que no ha encontrado respuesta normativa adecuada en la mayoría de los ordenamientos que lo utilizan.

En el ámbito del reconocimiento facial, el estudio de Buolamwini & Gebru (2020) sobre Gender Shades documentó tasas de error de hasta el 34% en la identificación de mujeres con piel oscura en los sistemas comerciales de reconocimiento facial más utilizados, frente a tasas de error inferiores al 1% en hombres con piel clara. El Instituto Nacional de Estándares y Tecnología de los Estados Unidos (NIST, 2019) confirmó estos hallazgos en su evaluación comparativa de 189 algoritmos de reconocimiento facial, encontrando tasas de falsos positivos entre 10 y 100 veces mayores en rostros afroamericanos e indígenas americanos que en rostros caucásicos. La implicación procesal penal es directa: si un sistema de reconocimiento facial identifica a una persona como sospechosa con una tasa de error dramáticamente más alta cuando esa persona es mujer y de piel oscura, su uso en investigaciones criminales sin controles adecuados viola sistemáticamente el derecho a la no discriminación (Garvie et al., 2022; Eubanks, 2018).

La respuesta normativa a esta problemática ha avanzado de manera significativamente asimétrica entre jurisdicciones. La Unión Europea ha adoptado la respuesta más restrictiva: el Reglamento de IA de 2024 prohíbe con carácter general la identificación biométrica remota en tiempo real en espacios públicos con fines de represión penal, con excepciones limitadas y sujetas a autorización judicial previa. Estados Unidos, en ausencia de regulación federal, ha experimentado un mosaico de respuestas: varias ciudades (San Francisco, Boston, Baltimore) y estados han prohibido el uso del reconocimiento facial por las fuerzas del orden, mientras

que otros estados no tienen restricciones. En América Latina, la ausencia de regulación específica es casi generalizada, con la excepción de las directrices de la Autoridad Nacional de Protección de Datos de Brasil (ANPD, 2023) y de algunas resoluciones de la Corte Suprema argentina sobre el uso de reconocimiento facial en el sistema de justicia (Palazzi, 2022; CIDH, 2021).

La dignidad humana y el principio de autonomía personal

El principio de dignidad humana, que Kant (1785) formuló como la exigencia de tratar a la humanidad, tanto en la propia persona como en la de cualquier otro, siempre como un fin y nunca únicamente como un medio, adquiere una dimensión específica y perturbadora en el contexto de la justicia penal automatizada. Cuando un sistema algorítmico decide que una persona debe permanecer en prisión preventiva porque su perfil estadístico corresponde al de personas que tienen una alta probabilidad de reincidir, está, en un sentido filosóficamente significativo, tratando a esa persona como el representante de un grupo estadístico antes que como un individuo con sus propias circunstancias, historia y capacidad de autodeterminación (Nussbaum, 2006; Rawls, 1971).

Esta instrumentalización estadística del individuo no solo viola el principio kantiano de dignidad, sino que entra en conflicto con la premisa filosófica que subyace a toda la arquitectura del derecho penal liberal: la idea de que el ser humano es un agente moral libre y responsable, capaz de determinarse a sí mismo más allá de sus condicionamientos biológicos, sociales y estadísticos. Si el sistema penal acepta que una persona puede ser privada de libertad preventivamente en función de características demográficas y estadísticas que comparte con grupos de alta reincidencia, está abandonando implícitamente el principio de culpabilidad individual que ha fundamentado el derecho penal moderno desde la Ilustración, y adoptando en su lugar una lógica neopeligrosista que la dogmática penal garantista ha rechazado con energía en todas sus manifestaciones históricas (Ferrajoli, 2011; Zaffaroni, 2019).

El Comité de Derechos Humanos de las Naciones Unidas, en su Observación General núm. 35 sobre el artículo 9 del PIDCP (libertad y seguridad personales), ha establecido que la detención preventiva solo puede justificarse por razones individuales y concretas, no por evaluaciones estadísticas de grupos. Esta posición fue reafirmada en las Directrices de Kampala sobre

Condiciones de la Detención (2020), que establecen explícitamente que las evaluaciones de riesgo algorítmico no pueden constituir el fundamento suficiente para privar de libertad a una persona. El Grupo de Trabajo sobre Detención Arbitraria de Naciones Unidas ha adoptado conclusiones similares en varios casos individuales en que sistemas actuariales de evaluación de riesgo fueron utilizados para justificar la prisión preventiva (CDH, 2023).

El derecho al debido proceso y a la tutela judicial efectiva

El derecho al debido proceso (due process of law), consagrado en el artículo 14 del PIDCP, en el artículo 6 del CEDH, en el artículo 8 de la Convención Americana sobre Derechos Humanos y en los textos constitucionales de la casi totalidad de los sistemas democráticos, es el derecho que más directamente estructura las exigencias jurídicas que la justicia penal automatizada debe satisfacer. Este derecho incluye, entre sus componentes esenciales relevantes para la IA: el derecho a ser juzgado por un tribunal independiente e imparcial; el derecho a conocer los cargos que se formulan y la evidencia en que se fundamentan; el derecho a una defensa efectiva; el derecho a que las decisiones sean motivadas; y el derecho a recurrir las decisiones ante una instancia superior. Cada uno de estos componentes se ve amenazado de manera específica por el uso de sistemas algorítmicos en el proceso penal.

El derecho a conocer la evidencia en que se fundamenta la acusación se ve amenazado cuando la evidencia algorítmica es opaca por naturaleza técnica o por protección de secreto comercial, y la defensa no puede acceder a la información necesaria para cuestionarla de manera efectiva. Como se analizó en el Capítulo IV, la invocación del secreto comercial por proveedores de sistemas de IA como COMPAS para impedir el acceso de la defensa al código fuente y a los datos de entrenamiento ha sido el eje de una jurisprudencia contradictoria en los Estados Unidos. El Tribunal Supremo de Wisconsin, en State v. Loomis (2016), avaló el uso de COMPAS en la fase de sentencia pese a que la defensa no podía acceder al algoritmo, argumentando que el juez no había basado su decisión exclusivamente en la valoración del sistema. Esta solución, criticada unánimemente por la doctrina especializada, es paradigmática de las tensiones irresueltas entre la admisión de evidencia algorítmica y las garantías del proceso equitativo (Završnik, 2021; Kerr, 2020).

El derecho a la motivación de las resoluciones judiciales plantea exigencias específicas ante la evidencia algorítmica que ningún ordenamiento ha resuelto plenamente. La motivación

judicial exige que el juez explique las razones por las que otorga valor probatorio a cada elemento de prueba. Cuando ese elemento de prueba es la conclusión de un sistema de IA que el juez no puede explicar técnicamente, la motivación formal corre el riesgo de convertirse en una ratificación acrítica de la conclusión algorítmica, privada del contenido de deliberación racional que justifica la autoridad de la decisión judicial (Hildebrandt, 2020; Pasquale, 2020). El Tribunal Europeo de Derechos Humanos, en su jurisprudencia sobre el artículo 6 del CEDH, ha establecido que la motivación judicial no necesita responder a todos los argumentos de las partes, pero sí debe abordar los que son determinantes para el resultado del proceso; cuando la conclusión algorítmica es determinante para el resultado, su aceptación acrítica no satisface esta exigencia (TEDH, Taxquet c. Bélgica, 2010).

El derecho de defensa efectiva se ve amenazado cuando existe una asimetría estructural entre los recursos técnicos disponibles para la acusación y para la defensa en el examen de la evidencia algorítmica. La acusación tiene habitualmente acceso privilegiado a los sistemas utilizados por las fuerzas del orden, mientras que la defensa depende de los recursos del acusado para contratar peritos especializados que puedan examinar e impugnar esas conclusiones. En los sistemas de defensa pública, cuya financiación es frecuentemente insuficiente en la mayoría de los países, esta asimetría puede ser tan pronunciada que comprometa la igualdad de armas que el proceso equitativo requiere (Ferguson, 2021; Završnik, 2021). La Comisión Interamericana de Derechos Humanos ha señalado que los Estados tienen la obligación de garantizar que los servicios de defensa pública dispongan de los recursos técnicos necesarios para impugnar la evidencia algorítmica cuando esta sea determinante para la decisión (CIDH, 2021).

El derecho a la presunción de inocencia

El principio de presunción de inocencia, consagrado en el artículo 11.1 de la Declaración Universal de Derechos Humanos, en el artículo 14.2 del PIDCP y en el artículo 6.2 del CEDH, es amenazado de manera específica por los sistemas de evaluación de riesgo actuarial que predicen la probabilidad futura de conducta delictiva para fundamentar decisiones restrictivas de derechos. La lógica de estos sistemas invierte la presunción de inocencia: en lugar de partir de la presunción de que el individuo no delinquirá, parten de la predicción estadística de que delinquirá y utilizan esa predicción para justificar su detención preventiva.

El Comité de Derechos Humanos de las Naciones Unidas ha establecido, en su Observación General núm. 32, que la presunción de inocencia impone a las autoridades públicas la obligación de no prejuzgar el resultado del juicio, y que toda medida privativa de libertad antes de la condena debe estar justificada por razones individuales y concretas de peligro de fuga, riesgo de destrucción de pruebas u obstaculización del proceso, o peligro para la seguridad pública. El uso de un sistema actuarial que predice la reincidencia en función de características demográficas y estadísticas del grupo de referencia para justificar la prisión preventiva no satisface el estándar de razones individuales y concretas que el Comité exige (CDH, 2023).

La compatibilidad de los sistemas actuariales de evaluación de riesgo con la presunción de inocencia es uno de los debates doctrinales más intensos en el campo. La posición dominante en la doctrina continental europea, representada por autores como Ferrajoli (2011), Zaffaroni (2019) y Hassemer (2018), es que el uso de evaluaciones actuariales para fundamentar la prisión preventiva viola la presunción de inocencia en su dimensión de garantía contra el derecho penal de autor, que condena por lo que se es o por lo que estadísticamente se es antes que por lo que se ha hecho. La posición dominante en la literatura angloamericana es más matizada, y autora como Hannah-Moffat (2019) arguye que la evaluación de riesgo puede ser compatible con la presunción de inocencia si se utiliza como herramienta de reducción del encarcelamiento preventivo (identificando a quienes no necesitan ser detenidos) antes que como justificación para incrementarlo.

El derecho a la privacidad y la protección de datos personales

El derecho a la privacidad, consagrado en el artículo 17 del PIDCP, en el artículo 8 del CEDH y en las constituciones de la gran mayoría de los Estados democráticos, adquiere una dimensión específica y particularmente amenazada en el contexto de la vigilancia algorítmica orientada a la prevención del delito. Los sistemas de policing predictivo, que analizan grandes volúmenes de datos sobre las actividades, movimientos, relaciones y comunicaciones de las personas para predecir dónde y cuándo es probable que se cometan delitos o quién es más probable que los cometa, operan en la intersección entre la prevención del delito y la vigilancia masiva de la población.

El Relator Especial de las Naciones Unidas sobre el derecho a la privacidad, en su informe de 2021, subrayó que los sistemas de vigilancia algorítmica orientados a la seguridad pública

constituyen una amenaza sistémica al derecho a la privacidad cuando operan sin los principios de necesidad, proporcionalidad, transparencia y control democrático. En particular, señaló que la construcción de perfiles individuales mediante el análisis automatizado de datos dispersos (el denominado profiling algorítmico) puede vulnerar el derecho a la privacidad incluso cuando los datos individuales utilizados son públicos o están lícitamente disponibles, porque su agregación y análisis produce una imagen íntima de la persona que esta no ha consentido revelar y que puede ser utilizada de manera que afecte significativamente sus derechos (Cannataci, 2021).

La Unión Europea ha desarrollado el marco más avanzado de protección de la privacidad en el contexto de la IA: el Reglamento General de Protección de Datos (RGPD, 2016/679) prohíbe las decisiones basadas exclusivamente en el tratamiento automatizado de datos personales que produzcan efectos jurídicos significativos o que afecten de manera igualmente significativa a las personas (artículo 22), con excepciones limitadas y sujetas a salvaguardas específicas. La Directiva de Protección de Datos en el Ámbito Policial (2016/680) establece requisitos específicos para el tratamiento de datos personales con fines de prevención, investigación, detección o enjuiciamiento de infracciones penales, incluyendo la exigencia de base legal explícita, proporcionalidad y supervisión independiente. En América Latina, el nivel de protección es significativamente menor: solo Brasil, Argentina, Colombia, Chile y Uruguay cuentan con legislación general de protección de datos personales, y ninguna de estas leyes aborda de manera específica el tratamiento algorítmico de datos con fines de seguridad pública (Palazzi, 2022; CIDH, 2021).

Derechos humanos afectados por la automatización de la justicia penal: análisis comparado

Derecho afectado	Instrumento internacional	Mecanismo de amenaza algorítmica	Respuesta normativa UE	Respuesta AL (evaluación)
No discriminación e igualdad	DUDH art. 7; PIDCP art. 26; CEDH art. 14	Sesgos en entrenamiento; tasas de error	Reglamento IA: prohibición de discriminación algorítmica;	Deficiente; sin regulación específica en la

Derecho afectado	Instrumento internacional	Mecanismo de amenaza algorítmica	Respuesta normativa UE	Respuesta AL (evaluación)
		diferenciadas por raza/género	auditorías de sesgo	mayoría de países
Dignidad humana	DUDH art. 1; PIDCP preámbulo; CEDH art. 3	Reducción del individuo a estadística grupal; lógica actuarial peligrosista	Reglamento IA: prohibición de sistemas de puntuación social; supervisión humana	Incipiente; referenciada en algunas políticas nacionales de IA sin efectos vinculantes
Debido proceso	PIDCP art. 14; CEDH art. 6; CADH art. 8	Opacidad; asimetría de armas; motivación insuficiente; secreto comercial	Reglamento IA: explicabilidad; derecho de impugnación; supervisión judicial	Escasa; solo Colombia y Chile tienen disposiciones generales sobre evidencia digital
Presunción de inocencia	DUDH art. 11.1; PIDCP art. 14.2; CEDH art. 6.2	Sistemas actuariales en prisión preventiva; lógica predictiva vs. presunción de inocencia	Reglamento IA: prohibición de determinar culpabilidad exclusivamente mediante IA	Muy deficiente; ausencia total de regulación específica en todos los países analizados

Derecho afectado	Instrumento internacional	Mecanismo de amenaza algorítmica	Respuesta normativa UE	Respuesta AL (evaluación)
Privacidad y protección de datos	PIDCP art. 17; CEDH art. 8; CADH art. 11	Vigilancia masiva; profiling algorítmico; reconocimiento facial	RGPD art. 22; Directiva 2016/680; Reglamento IA: restricciones biométricas	Parcial; leyes de protección de datos en Brasil, Argentina, Colombia, Chile, Uruguay; sin aplicación específica a IA penal

Nota: Elaboración propia a partir de CDH (2023); Cannataci (2021); CIDH (2021); Comisión Europea (2024); Palazzi (2022).

Principios de diseño ético de la IA para la justicia penal

El principio de supervisión humana significativa

El principio de supervisión humana significativa (meaningful human oversight) es el principio ético más universalmente consagrado en los marcos regulatorios de la IA aplicada a la justicia penal, y también el más frecuentemente vaciado de contenido en la práctica. La distinción entre supervisión humana significativa y supervisión humana puramente formal es crucial: la primera implica que el ser humano que supervisa el sistema tiene la información, la formación y el tiempo necesarios para comprender las conclusiones del sistema y para cuestionarlas cuando corresponda; la segunda se limita a la presencia nominal de un humano en el proceso de decisión que en la práctica ratifica automáticamente las conclusiones del sistema (Cummings, 2017; Floridi et al., 2022).

La investigación empírica sobre el comportamiento de los jueces ante las recomendaciones de sistemas de evaluación de riesgo algorítmico ha documentado un fenómeno preocupante: el efecto de automaticidad, por el cual los juzgadores tienden a seguir las recomendaciones del sistema con una frecuencia muy superior a la que correspondería si realizaran una

evaluación independiente de los datos subyacentes. Dressel & Farid (2022) documentaron que los jueces que recibieron la evaluación de riesgo de COMPAS siguieron sus recomendaciones en más del 78% de los casos, aun cuando la precisión del sistema no era superior a la del juicio de personas sin formación jurídica. Este efecto de automaticidad convierte la supervisión humana formal en una ratificación cuasi-automática de la conclusión algorítmica, privada del contenido deliberativo que la justifica (Cummings, 2017; Završnik, 2021).

Para que la supervisión humana sea genuinamente significativa en el ámbito de la justicia penal, la literatura especializada identifica al menos tres condiciones necesarias. La primera es la formación adecuada: los juzgadores y operadores jurídicos que interactúan con sistemas de IA deben recibir formación específica sobre sus capacidades, limitaciones y tasas de error, que les permita ejercer un juicio crítico independiente sobre sus conclusiones. La segunda es la disponibilidad de tiempo suficiente: la presión de las cargas de trabajo de los tribunales de instrucción es uno de los factores que más favorece la automaticidad en el seguimiento de las recomendaciones algorítmicas; la reducción del número de casos gestionados por cada juzgador, aunque costosa, es una condición estructural de la supervisión significativa. La tercera es la transparencia de la información: el juzgador debe tener acceso no solo a la conclusión del sistema sino a los datos subyacentes y a los factores que determinaron esa conclusión, para poder verificar independientemente si son correctos y pertinentes en el caso concreto (Cummings, 2017; Dressel & Farid, 2022; Floridi et al., 2022).

El principio de no delegación de decisiones irreversibles

El principio de no delegación de decisiones irreversibles (también formulado como principio de irrenunciabilidad de la decisión jurisdiccional humana) establece que determinadas categorías de decisiones, por su gravedad, su irreversibilidad o su incidencia directa sobre los derechos fundamentales de las personas, no pueden ser delegadas a sistemas algorítmicos, con independencia del grado de precisión que estos alcancen. Este principio, formulado con especial claridad por el Consejo de Europa en su Recomendación CM/Rec(2020)1 y recogido en el Reglamento de IA de la UE (2024/1689), tiene aplicaciones especialmente relevantes en el ámbito penal.

Las categorías de decisiones que la doctrina considera no delegables a sistemas algorítmicos en el ámbito penal incluyen: la determinación de la culpabilidad del acusado; la imposición de la pena; la decisión sobre la prisión preventiva cuando esta se funda en la valoración individual de las circunstancias del caso; y cualquier decisión que implique la valoración de circunstancias atenuantes o agravantes de carácter individualizado. En todas estas categorías, el elemento de valoración judicial independiente de las circunstancias concretas del caso es constitutivo de la legitimidad de la decisión, y su sustitución por una evaluación estadística grupal viola el principio de individualización de la respuesta penal que el derecho penal liberal consagra (Ferrajoli, 2011; Zaffaroni, 2019; Hildebrandt, 2020).

El Reglamento de IA de la UE (2024/1689) materializa este principio al establecer explícitamente, en su artículo 22, que los sistemas de IA no pueden utilizarse por las autoridades judiciales para adoptar decisiones en materia penal sin supervisión humana adecuada, y que la determinación de la culpabilidad de una persona no puede basarse exclusivamente en el resultado de un sistema de IA. El Tribunal de Justicia de la Unión Europea, en su interpretación del artículo 22 del RGPD, ha establecido además que el derecho a no ser objeto de decisiones basadas exclusivamente en el tratamiento automatizado se aplica a todas las decisiones que produzcan efectos jurídicos significativos o que afecten de manera igualmente significativa a las personas, categoría que comprende sin duda las decisiones penales de mayor impacto (TJUE, C-634/21, SCHUFA Holding AG, 2023).

El principio de equidad algorítmica y no discriminación por diseño

El principio de equidad algorítmica (algorithmic fairness) ha generado una de las controversias técnicas y filosóficas más fructíferas en la literatura de inteligencia artificial y derecho. El debate se articula en torno a la existencia de múltiples definiciones matemáticamente incompatibles de equidad que cualquier sistema de IA puede optimizar, pero no simultáneamente. Las tres definiciones más relevantes para la justicia penal son: la equidad demográfica (demographic parity), que exige que el sistema produzca resultados similares para grupos definidos por características protegidas; la equidad predictiva (predictive parity), que exige que la tasa de precisión del sistema sea similar para todos los grupos; y la equidad de calibración, que exige que las probabilidades predichas correspondan a las frecuencias observadas de manera similar para todos los grupos.

Chouldechova (2017) y Kleinberg et al. (2018) demostraron matemáticamente que estas tres definiciones son simultáneamente satisfacibles solo en condiciones muy restrictivas que no se dan en la práctica de la justicia penal, donde las tasas de delito medidas son sustancialmente diferentes entre grupos. Esta imposibilidad matemática tiene consecuencias normativas de primera magnitud: los diseñadores de sistemas de IA para la justicia penal deben necesariamente elegir qué tipo de equidad priorizan, y esa elección tiene consecuencias distributivas concretas sobre qué grupos son perjudicados por los errores del sistema. Esta elección no es técnica sino política y ética, y su adopción no puede ser dejada exclusivamente a los ingenieros y las empresas proveedoras del sistema.

La implicación normativa de este análisis es que el diseño de sistemas de IA para la justicia penal debe incluir un proceso deliberativo explícito sobre el tipo de equidad que el sistema debe optimizar, con participación de todos los grupos afectados, incluyendo especialmente aquellos que históricamente han sido más afectados por las desigualdades del sistema de justicia. Este proceso deliberativo es un requisito de legitimidad democrática antes que una exigencia técnica, y su ausencia en los procesos de contratación y despliegue de sistemas algorítmicos en el ámbito de la justicia penal es uno de los principales déficits de gobernanza que la literatura ha identificado (O'Neil, 2016; Eubanks, 2018; Richardson et al., 2019).

El principio de rendición de cuentas algorítmica (algorithmic accountability)

El principio de rendición de cuentas algorítmica exige que existan mecanismos efectivos para determinar quién es responsable cuando un sistema de IA produce resultados incorrectos, discriminatorios o dañinos en el ámbito de la justicia penal, y para garantizar que esa responsabilidad genere consecuencias que disuadan la repetición del daño e incentiven la adopción de mejores prácticas. Como se analizó en el Capítulo II de esta obra, los mecanismos de responsabilidad penal son insuficientes para cubrir el vacío de rendición de cuentas que los sistemas autónomos de IA generan. El principio de rendición de cuentas algorítmica plantea la necesidad de mecanismos complementarios de naturaleza regulatoria, técnica y procedimental que cierren ese vacío.

La literatura especializada identifica cuatro dimensiones de la rendición de cuentas algorítmica que son relevantes para la justicia penal. La primera es la rendición de cuentas por el diseño (design accountability): los desarrolladores y fabricantes de sistemas de IA para la

justicia penal deben ser responsables de la calidad, fiabilidad y ausencia de sesgos discriminatorios de sus sistemas, y deben poder demostrar mediante documentación técnica certificada que han adoptado todas las medidas necesarias para minimizar esos riesgos. La segunda es la rendición de cuentas por el despliegue (deployment accountability): las autoridades que deciden contratar y desplegar sistemas de IA en el ámbito de la justicia penal deben ser responsables de verificar que los sistemas cumplen con los estándares técnicos y éticos aplicables, y de establecer protocolos de supervisión y corrección durante su operación. La tercera es la rendición de cuentas por el uso (use accountability): los operadores individuales que utilizan los sistemas de IA en sus funciones jurisdiccionales o de investigación deben ser responsables de ejercer una supervisión crítica de las conclusiones del sistema, y de no delegarle decisiones que son de su responsabilidad personal. La cuarta es la rendición de cuentas institucional (institutional accountability): las instituciones del sistema de justicia en su conjunto deben establecer mecanismos de revisión periódica del impacto de los sistemas de IA sobre los derechos de las personas afectadas, incluyendo especialmente el análisis del impacto diferencial sobre grupos vulnerables (Mittelstadt et al., 2016; Floridi et al., 2022; Wachter et al., 2020).

Gobernanza ético-jurídica de la IA penal: modelos comparados

El modelo europeo: regulación vinculante ex ante

El modelo europeo de gobernanza de la IA en el ámbito penal es el más avanzado en términos de vinculatoriedad jurídica y de cobertura normativa. El Reglamento de IA (2024/1689) articula un sistema de gobernanza basado en la clasificación de riesgos que establece requisitos diferenciados según el nivel de riesgo que el sistema presenta para los derechos fundamentales. Para los sistemas de alto riesgo utilizados en el ámbito de la justicia penal, el Reglamento impone: obligaciones de evaluación de conformidad antes del despliegue; documentación técnica completa; registro en una base de datos pública de la UE; sistemas de gestión de riesgos continuos durante el ciclo de vida del sistema; transparencia e información a los afectados; supervisión humana adecuada; y derechos de impugnación de las decisiones fundamentadas en sus conclusiones. El cumplimiento de estos requisitos está supervisado por autoridades nacionales de supervisión de mercado designadas por cada Estado miembro, coordinadas a nivel europeo por la Oficina Europea de Inteligencia Artificial (EUAI Office).

La Directiva sobre el uso de inteligencia artificial en el sistema judicial (en proceso de elaboración al cierre del corpus) está llamada a completar este marco con disposiciones procesales penales específicas sobre la admisibilidad de la evidencia algorítmica, los requisitos de motivación de las decisiones judiciales fundadas en sistemas de IA y los mecanismos de auditoría independiente. La Recomendación CM/Rec(2020)1 del Consejo de Europa, aunque de carácter no vinculante, ha servido como referencia normativa para los tribunales y legisladores de los 46 Estados miembros del Consejo, orientando la interpretación de las garantías del artículo 6 del CEDH en el contexto de la evidencia y las decisiones algorítmicas (Consejo de Europa, 2020; Comisión Europea, 2024).

Las fortalezas del modelo europeo son su carácter vinculante, su cobertura sistemática del ciclo de vida de los sistemas de IA, y la creación de mecanismos de supervisión institucional con capacidad sancionatoria (multas de hasta 35 millones de euros o el 7% del volumen de negocio global para las infracciones más graves). Sus limitaciones son la complejidad de implementación para organizaciones de menor tamaño, el riesgo de que los requisitos formales de documentación sustituyan a los mecanismos sustantivos de control ético, y la velocidad de evolución tecnológica, que puede hacer que los estándares del Reglamento queden obsoletos antes de que hayan sido plenamente implementados (Floridi et al., 2022; Hildebrandt, 2020).

El modelo estadounidense: regulación sectorial y fragmentada

El modelo estadounidense de gobernanza de la IA en el ámbito penal se caracteriza por la fragmentación normativa entre el nivel federal y los estados, por la prevalencia de la autorregulación y las directrices voluntarias sobre la regulación vinculante, y por la primacía del enfoque sectorial (por dominio de aplicación) sobre el enfoque sistémico. La Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (EO 14110, 2023) de la administración Biden estableció un marco de principios y exigencias de transparencia para los sistemas de IA de mayor impacto, pero fue significativamente debilitada por la administración Trump en 2025, que revocó varias de sus disposiciones mediante la Executive Order on Removing Barriers to American Leadership in Artificial Intelligence (EO 14179, 2025).

En el ámbito específico de la justicia penal, los Estados Unidos han adoptado un enfoque de regulación específica por modalidad. Varios estados y ciudades han adoptado moratorias o prohibiciones del uso del reconocimiento facial por las fuerzas del orden (Illinois Biometric Information Privacy Act; prohibiciones municipales en San Francisco, Boston, Portland, Denver). El DEFIANCE Act (2024) criminalizó a nivel federal la producción y distribución de deepfakes sexuales no consentidos. La Algorithmic Accountability Act, presentada en varias legislaturas del Congreso pero sin llegar a aprobarse, habría establecido requisitos de evaluación de impacto para los sistemas de decisión automatizada de alto impacto. El Blueprint for an AI Bill of Rights (2022), un documento de la Oficina de Política Científica y Tecnológica de la Casa Blanca, establece principios no vinculantes de protección frente a sistemas algorítmicos inseguros o ineficaces, diseño discriminatorio y vigilancia. Su carácter no vinculante limita significativamente su eficacia como instrumento de gobernanza (Department of Justice, 2023; Ferguson, 2021).

El modelo latinoamericano: desafíos y oportunidades emergentes

La situación de la gobernanza de la IA en el ámbito penal en América Latina presenta un panorama de notable heterogeneidad y de oportunidades aún insuficientemente aprovechadas. A diferencia de lo ocurrido en Europa, donde el Reglamento de IA ha establecido un marco vinculante regional, en América Latina la gobernanza de la IA se articula a través de políticas nacionales, estrategias de datos y, en algunos casos, proyectos de ley general de IA que se encuentran en distintas fases de elaboración (Palazzi, 2022; CIDH, 2021; Ortega & Salinas, 2023).

Brasil es el país latinoamericano con el marco más avanzado. El Proyecto de Ley 2338/2023, aprobado por el Senado y en tramitación en la Cámara de Diputados al cierre del corpus, establece un sistema de gobernanza de la IA basado en la evaluación de riesgos que tiene influencias explícitas del modelo europeo. En lo que respecta al ámbito penal, el proyecto clasifica como de alto riesgo los sistemas de IA utilizados para la administración de justicia y establece requisitos de transparencia, supervisión humana y derecho de recurso. La Autoridad Nacional de Protección de Datos (ANPD) ha publicado directrices sobre el uso del reconocimiento facial en espacios públicos y en el sistema de seguridad pública que representan un avance significativo. Colombia ha avanzado en la adopción de un marco de ética de datos y gobernanza algorítmica a través del documento CONPES 3988 (2021), que

aunque no es vinculante en sentido estricto establece principios orientadores relevantes. Chile, con su Política Nacional de Inteligencia Artificial (2021) y el proceso de elaboración de un proyecto de ley regulador de la IA, está siguiendo un camino similar. Argentina ha publicado lineamientos para una estrategia de IA que incluyen referencias a los principios de transparencia, explicabilidad y no discriminación, pero sin efectos normativos vinculantes. Ecuador, México, Perú y los países centroamericanos se encuentran en fases más incipientes de este proceso (Palazzi, 2022; Ortega & Salinas, 2023; CIDH, 2021).

La Comisión Interamericana de Derechos Humanos, en su Informe Temático sobre Tecnología y Derechos Humanos (CIDH, 2021), subrayó que los Estados de la región tienen obligaciones positivas derivadas de la Convención Americana de Derechos Humanos para garantizar que el uso de tecnologías de IA en el sistema de justicia no vulnere los derechos de las personas, especialmente de los grupos en situación de vulnerabilidad. El informe identificó como áreas de preocupación prioritaria el uso del reconocimiento facial en el sistema de seguridad pública, el uso de sistemas de evaluación de riesgo actuarial en decisiones de privación de libertad, y la ausencia de marcos regulatorios adecuados que garanticen el derecho de las personas afectadas a conocer que han sido objeto de una decisión algorítmica y a impugnarla efectivamente.

La Corte Interamericana de Derechos Humanos no se ha pronunciado aún específicamente sobre la compatibilidad de los sistemas de decisión algorítmica en el ámbito penal con la Convención Americana, pero la aplicación de su jurisprudencia sobre el artículo 8 (garantías judiciales) y el artículo 25 (protección judicial) al contexto algorítmico apunta hacia la obligación de los Estados de garantizar el acceso a recursos judiciales efectivos frente a decisiones automatizadas que afecten los derechos de las personas (CADH, arts. 8 y 25; Corte IDH, caso Escher y otros v. Brasil, 2009).

Modelos emergentes de auditoría ética algorítmica

Un desarrollo de notable interés en la literatura más reciente es el surgimiento de modelos de auditoría ética algorítmica como mecanismo de gobernanza intermedio entre la regulación vinculante ex ante y la responsabilidad ex post por daños. La auditoría ética algorítmica (AEA) es un proceso sistemático de evaluación del diseño, funcionamiento e impacto de un sistema de IA con respecto a un conjunto de estándares éticos y de derechos humanos predefinidos,

llevado a cabo por un organismo independiente del proveedor y del desplegador del sistema (Raji et al., 2020; Mokander & Floridi, 2021).

En el ámbito de la justicia penal, los modelos de AEA más desarrollados comprenden al menos las siguientes dimensiones de evaluación: (1) auditoría de datos: verificación de la calidad, representatividad y ausencia de sesgos en los datos de entrenamiento del sistema; (2) auditoría de sesgo: análisis de las tasas de error diferenciadas por grupos definidos por características protegidas y evaluación de su compatibilidad con los estándares de equidad aplicables; (3) auditoría de explicabilidad: verificación de la capacidad del sistema de proporcionar explicaciones comprensibles y suficientemente detalladas de sus conclusiones para permitir la contradicción procesal; (4) auditoría de seguridad: evaluación de la robustez del sistema frente a ataques adversariales y manipulaciones; y (5) auditoría de impacto en derechos humanos: evaluación del impacto real del despliegue del sistema sobre los derechos de las personas afectadas, con especial atención a grupos vulnerables (Raji et al., 2020; Mokander & Floridi, 2021; Metcalf & Moss, 2019).

Los modelos de AEA enfrentan obstáculos significativos para su implementación efectiva en el ámbito de la justicia penal. El principal es el acceso a la información necesaria para realizarla: los proveedores de sistemas propietarios típicamente se niegan a proporcionar el código fuente y los datos de entrenamiento necesarios para una auditoría completa, invocando el secreto comercial. Las soluciones propuestas incluyen la creación de organismos de auditoría acreditados con poderes de acceso garantizados legalmente, la exigencia de auditorías previas al despliegue como requisito de contratación pública, y el desarrollo de técnicas de auditoría de caja negra que permiten evaluar el comportamiento del sistema sin acceso directo a su arquitectura interna (Raji et al., 2020; Metcalf & Moss, 2019). El Reglamento de IA de la UE ha dado un paso en esta dirección al establecer que las autoridades nacionales de supervisión tendrán acceso a la documentación técnica completa de los sistemas de alto riesgo, y que los proveedores deberán someterse a evaluaciones de conformidad por parte de organismos de tercera parte acreditados para determinadas categorías de sistemas.

Modelos de gobernanza de la IA penal: análisis comparado

Dimensión	Modelo UE (Reglamento IA 2024)	Modelo EE.UU. (sectorial/voluntario)	Modelo latinoamericano (emergente)	Evaluación comparada
Carácter normativo	Vinculante; aplicación directa	Fragmentado; en mayoría voluntario	Principalmente no vinculante; en transición	UE: más robusto. EE.UU.: regresión 2025. AL: desarrollo urgente
Enfoque	Sistémico; basado en riesgos	Sectorial; basado en casos	Mixto; principios generales + casos específicos	El enfoque sistémico ofrece mayor cobertura y predictibilidad
Supervisión institucional	EUAI Office + autoridades nacionales; poderes sancionatorios	FTC; NIST; no hay autoridad específica de IA penal	Agencias de protección de datos + ministerios de justicia; capacidad técnica limitada	UE: más desarrollado. AL necesita refuerzo de capacidades técnicas
Derechos de las personas afectadas	Derecho a explicación; derecho de impugnación; revisión	Variable; depende del estado y del sector	Incompleto; escasa regulación específica en la mayoría de países	Brecha significativa en AL respecto a estándares internacionales de DDHH

Dimensión	Modelo UE (Reglamento IA 2024)	Modelo EE.UU. (sectorial/voluntario)	Modelo latinoamericano (emergente)	Evaluación comparada
	humana garantizada			
Auditoría algorítmica	Obligatoria para sistemas de alto riesgo; organismos acreditados	Voluntaria; algunos requisitos sectoriales (finanzas, salud)	Ausente o incipiente en todos los países de la región	Necesidad urgente de creación de organismos de auditoría independientes en AL

Nota: Elaboración propia a partir de Comisión Europea (2024); Department of Justice (2023); Ortega & Salinas (2023); CIDH (2021); Palazzi (2022).

El problema del sesgo sistémico y los datos contaminados

Uno de los hallazgos más consistentes y perturbadores de la revisión bibliográfica es la demostración empírica de que el sesgo algorítmico en los sistemas de justicia penal no es un defecto técnico accidental que puede corregirse con mejoras de ingeniería, sino la expresión matemática de inequidades históricas profundamente arraigadas en los datos sobre los que los sistemas son entrenados. Esta tesis, formulada con especial claridad por Richardson et al. (2019) en su trabajo seminal sobre "Dirty Data, Bad Predictions", tiene implicaciones de primera magnitud para la ética y la gobernanza de la IA penal.

Richardson et al. (2019) analizaron los sistemas de policing predictivo de varias ciudades de los Estados Unidos y encontraron que en todos los casos examinados, los datos históricos de delito sobre los que fueron entrenados habían sido producidos durante períodos en que las fuerzas del orden exhibían prácticas documentadas de patrullaje discriminatorio, registros motivados racialmente y violaciones de derechos civiles. Los sistemas entrenados sobre estos datos históricos "aprenden" a predecir los delitos donde la policía ha históricamente buscado

con mayor intensidad, lo que produce un sesgo circular de retroalimentación: el sistema dirige más recursos policiales a los barrios históricamente sobrepatrullados, generando más detecciones en esas zonas, que a su vez refuerzan las predicciones del sistema. Este círculo vicioso algorítmico convierte el sesgo histórico en una predicción matemáticamente fundada de la realidad presente, confiriéndole una apariencia de legitimidad técnica que lo hace más difícil de combatir que el sesgo humano abiertamente visible.

Eubanks (2018), en su exhaustivo análisis del uso de sistemas algorítmicos de decisión en los servicios sociales de los Estados Unidos, documentó un fenómeno análogo que denominó automatización de la desigualdad: los sistemas de decisión automatizada que determinan si una persona tiene acceso a beneficios sociales, atención sanitaria o protección de la infancia reproducen sistemáticamente las inequidades del sistema que los genera, operando como una forma de discriminación estructural que se beneficia de la apariencia de neutralidad matemática para evadir el escrutinio que se aplicaría a decisiones humanas igualmente sesgadas. La trasposición de este análisis al ámbito de la justicia penal es directa: los sistemas de evaluación de riesgo actuarial que determinan si una persona permanece en prisión preventiva o accede a la libertad condicional reproducen y amplían las desigualdades históricas del sistema penal, especialmente las que afectan a las personas de color y a las personas en situación de pobreza.

La respuesta normativa a este problema requiere ir más allá de las exigencias técnicas de desbiasado (debiasing) de los sistemas algorítmicos: requiere un cuestionamiento de la calidad ética de los datos históricos que se utilizan para entrenarlos, y en algunos casos la prohibición del uso de determinadas categorías de datos como base para sistemas de decisión que afectan a derechos fundamentales. El Reglamento de IA de la UE aborda parcialmente este problema al exigir que los sistemas de alto riesgo sean entrenados sobre conjuntos de datos que sean pertinentes, representativos y suficientemente completos para minimizar los riesgos de sesgos discriminatorios. Sin embargo, la evaluación de la calidad de los datos de entrenamiento en función de sus posibles sesgos históricos requiere competencias técnicas y normativas específicas que los organismos de supervisión actuales no siempre poseen (Richardson et al., 2019; Eubanks, 2018; O'Neil, 2016).

Hacia un estatuto ético-jurídico de la inteligencia artificial en la justicia penal

La síntesis de los resultados del análisis permite articular los elementos constitutivos de un estatuto ético-jurídico de la inteligencia artificial en la justicia penal, que vaya más allá de las prohibiciones y restricciones sectoriales para establecer un marco coherente de principios, derechos, obligaciones y mecanismos de garantía. Este estatuto, que aquí se propone como horizonte normativo antes que como descripción del derecho vigente, se articula en torno a cinco ejes.

El primer eje es el catálogo de usos prohibidos, que comprende aquellas aplicaciones de la IA en el ámbito de la justicia penal que son incompatibles con los principios éticos y los derechos humanos con independencia de su nivel de precisión técnica. Este catálogo debe incluir: la determinación de la culpabilidad penal basada exclusivamente en conclusiones algorítmicas; la imposición de la pena sin valoración judicial individualizada; el uso de sistemas de puntuación social con efectos en el acceso a la justicia; el despliegue de sistemas de identificación biométrica masiva en espacios públicos sin autorización judicial previa y proporcional; y el uso de sistemas de IA cuyos sesgos discriminatorios han sido documentados sin haber sido corregidos.

El segundo eje es el catálogo de usos condicionados, que comprende las aplicaciones de la IA en la justicia penal que pueden ser eticamente admisibles bajo condiciones específicas de diseño, transparencia, supervisión y control. Este catálogo incluye los sistemas de apoyo a la investigación criminal, los sistemas de evaluación de riesgo como herramienta de orientación no determinante para el juez, los sistemas de análisis forense digital, y los sistemas de procesamiento de grandes volúmenes de evidencia. Para cada uno de estos usos, el estatuto debe especificar las condiciones éticas de admisibilidad: ausencia de sesgo discriminatorio documentado, explicabilidad mínima de las conclusiones, supervisión humana significativa, derecho de impugnación de la parte afectada, y auditoría periódica independiente.

El tercer eje es el conjunto de derechos de las personas afectadas por decisiones algorítmicas en el ámbito penal. Estos derechos deben incluir: el derecho a ser informado de que una decisión que le afecta ha sido adoptada con apoyo determinante de un sistema algorítmico; el derecho a conocer el tipo de sistema utilizado y los datos que el sistema consideró relevantes; el derecho a impugnar la conclusión del sistema ante un órgano humano

competente; el derecho a que la decisión sea revisada por un ser humano que pueda valorar las circunstancias individuales del caso; y el derecho a reparación cuando una decisión algorítmica discriminatoria o incorrecta haya producido un daño.

El cuarto eje es el conjunto de obligaciones de los Estados en materia de gobernanza de la IA penal. Estas obligaciones incluyen: la obligación de no desplegar sistemas de IA en el ámbito de la justicia penal que no hayan superado una evaluación de impacto sobre los derechos humanos; la obligación de garantizar la formación técnica adecuada de los operadores del sistema de justicia que interactúan con sistemas de IA; la obligación de crear y financiar organismos de auditoría independiente con los recursos técnicos y jurídicos necesarios; la obligación de revisar periódicamente el impacto de los sistemas desplegados sobre los derechos de las personas afectadas; y la obligación de reparar los daños causados por sistemas de IA defectuosos o discriminatorios.

El quinto eje es el conjunto de mecanismos de rendición de cuentas algorítmica, que articulan la responsabilidad de los distintos actores del ecosistema de la IA penal. Estos mecanismos deben incluir la responsabilidad administrativa de los proveedores por el incumplimiento de los requisitos de diseño ético; la responsabilidad penal de las personas físicas que adopten decisiones que violen los derechos de las personas afectadas utilizando sistemas de IA como pretexto de objetividad; la responsabilidad de las instituciones por los daños causados por sistemas que han sido desplegados sin las garantías necesarias; y mecanismos de compensación colectiva para los grupos sistémicamente perjudicados por sesgos algorítmicos que han operado durante períodos prolongados.

DISCUSIÓN Y CONCLUSIONES

Discusión de los hallazgos

Los resultados del análisis desarrollado en el presente capítulo permiten formular un conjunto de reflexiones críticas que sitúan los hallazgos en el contexto más amplio del debate sobre la ética de la tecnología, los derechos humanos y el futuro de la justicia penal. Una primera reflexión versa sobre la relación entre los marcos éticos internacionales y su implementación efectiva. La proliferación de declaraciones de principios, recomendaciones y marcos éticos voluntarios sobre la IA en los últimos años ha creado lo que Hagendorff (2020) denominó un problema de brecha de implementación: la distancia entre los principios abstractamente

proclamados y los mecanismos concretos de su aplicación es tan grande que los marcos éticos se convierten, en la práctica, en instrumentos de lavado ético (ethics washing) que otorgan a los proveedores de sistemas de IA una apariencia de responsabilidad sin los costes de una rendición de cuentas genuina. La experiencia del Reglamento de IA de la UE, el primer instrumento vinculante de alcance regional, demostrará en los próximos años si es posible cerrar esa brecha mediante la regulación vinculante, o si las dificultades de implementación la perpetúan bajo formas más sofisticadas.

Una segunda reflexión concierne a la tensión irresoluble entre la promesa de objetividad del algoritmo y la naturaleza fundamentalmente normativa y política de las decisiones de justicia penal. La ilusión de objetividad algorítmica es, quizás, el mayor riesgo ético que plantea la automatización de la justicia penal: no porque los sistemas de IA sean inherentemente maliciosos, sino porque su aparente neutralidad matemática puede ocultar y legitimar elecciones profundamente políticas sobre qué bienes se protegen, qué conductas se castigan y qué grupos se vigilan. La pregunta '¿qué predice el algoritmo?' siempre presupone la pregunta '¿qué queremos optimizar?', y esta segunda pregunta es irreductiblemente normativa. La respuesta que los diseñadores de sistemas de IA dan implícitamente a esta pregunta en sus elecciones de diseño tiene consecuencias distributivas concretas que afectan a personas reales, y merece el mismo escrutinio democrático que cualquier otra decisión de política criminal (O'Neil, 2016; Richardson et al., 2019; Pasquale, 2020).

Una tercera reflexión se refiere a la especificidad del contexto latinoamericano. Los análisis desarrollados en la literatura especializada, en su gran mayoría elaborados en el contexto de las democracias avanzadas del Norte Global, plantean problemas de transferibilidad que la investigación regional debe abordar. En América Latina, la vulnerabilidad específica ante los riesgos de la IA penal está agravada por tres factores: la mayor desigualdad estructural y los sesgos históricos más pronunciados de los sistemas de justicia penal de la región, que hacen que los datos disponibles para entrenar sistemas de IA sean más sesgados que en otros contextos; la menor capacidad institucional de los organismos de supervisión y auditoría, que hace más difícil el control efectivo de los sistemas desplegados; y la mayor vulnerabilidad de los grupos más afectados por el sesgo algorítmico (comunidades indígenas, afrodescendientes, personas en situación de pobreza) que tienen menor acceso a los

mecanismos de impugnación y reparación disponibles (Zaffaroni, 2019; CIDH, 2021; Palazzi, 2022).

Los hallazgos del presente capítulo se articulan coherentemente con los análisis de los capítulos precedentes de esta obra. La crisis de imputación penal identificada en el Capítulo II encuentra en el presente capítulo su correlato ético: el vacío de responsabilidad jurídica es también un vacío de rendición de cuentas ética. Las nuevas tipologías delictivas analizadas en el Capítulo III, especialmente la desinformación algorítmica y la manipulación de sistemas de decisión automatizada, adquieren en el contexto del presente capítulo una dimensión adicional: son amenazas no solo a bienes jurídicos individuales, sino a los fundamentos éticos de la deliberación democrática. Y la problemática de la evidencia algorítmica examinada en el Capítulo IV se revela, en la perspectiva ética del presente capítulo, como una amenaza al valor epistémico y a la legitimidad moral del juicio penal como institución democrática.

En relación con la literatura internacional, los hallazgos del presente capítulo confirman y desarrollan las conclusiones de Završnik (2021) sobre la erosión de garantías procesales por la automatización de la justicia, de Hildebrandt (2020) sobre la tensión entre el rule of law y la tecnología algorítmica, y de Floridi et al. (2022) sobre la necesidad de marcos éticos vinculantes para la IA. Al mismo tiempo, la presente revisión aporta elementos adicionales derivados de la incorporación de la perspectiva latinoamericana, de los instrumentos interamericanos de derechos humanos y de los desarrollos normativos más recientes, incluyendo el Reglamento de IA de la UE de 2024, el PL 2338/2023 de Brasil y los informes del ACNUDH y la CIDH de 2021 y 2023.

Las limitaciones del presente análisis son las inherentes a cualquier revisión bibliográfica de un campo en acelerada evolución: la posibilidad de que algunos desarrollos normativos o tecnológicos significativos hayan ocurrido tras el cierre del corpus; el sesgo geográfico de la literatura académica disponible, que sobrerepresenta las perspectivas del Norte Global; y la naturaleza prescriptiva de las propuestas formuladas, que son orientaciones normativas antes que descripción del derecho positivo vigente. La investigación futura deberá complementar este análisis con estudios empíricos sobre la implementación efectiva de los marcos de gobernanza en distintos contextos, con análisis de la jurisprudencia de los tribunales interamericanos sobre IA y derechos humanos a medida que esta se vaya desarrollando, y con evaluaciones del impacto real del Reglamento de IA de la UE en el ámbito de la justicia penal.

Conclusiones

El análisis efectuado en el presente capítulo permite formular las siguientes conclusiones, que complementan y proyectan sobre el plano ético-normativo los hallazgos de los capítulos precedentes de esta obra:

La intersección entre la ética de la inteligencia artificial y la justicia penal revela que la automatización del sistema punitivo no es éticamente neutra: implica elecciones de diseño que tienen consecuencias distributivas concretas sobre qué personas y grupos son más afectados por los errores y sesgos del sistema, y exige por tanto un nivel de escrutinio democrático y de rendición de cuentas equivalente al que se aplica a cualquier otra decisión de política criminal.

Los marcos éticos internacionales, aunque convergentes en sus principios fundamentales (supervisión humana, transparencia, equidad, rendición de cuentas), presentan una brecha de implementación significativa que solo la regulación vinculante, como la adoptada por la Unión Europea con el Reglamento de IA de 2024, puede aspirar a cerrar. Los marcos voluntarios y los principios sin mecanismos de cumplimiento son instrumentos insuficientes para garantizar la compatibilidad de la justicia penal automatizada con los derechos humanos.

Los derechos a la no discriminación, la dignidad humana, el debido proceso, la presunción de inocencia y la privacidad son sistemáticamente amenazados por el uso no controlado de sistemas algorítmicos en la justicia penal. La evidencia empírica sobre los sesgos de los sistemas de evaluación de riesgo y de reconocimiento facial demuestra que estas amenazas no son hipotéticas sino actuales y documentadas, con consecuencias reales sobre vidas humanas concretas.

El principio de supervisión humana significativa es la garantía más fundamental de la compatibilidad entre la IA y el Estado de Derecho en el ámbito penal, pero su implementación efectiva requiere condiciones estructurales (formación adecuada, tiempo suficiente, acceso a información) que los sistemas de justicia actuales frecuentemente no proveen. La reducción de la supervisión humana a una ratificación formal de conclusiones algorítmicas vacía de contenido esta garantía esencial.

El sesgo algorítmico en los sistemas de justicia penal es, en la mayoría de los casos documentados, la expresión matemática de inequidades históricas profundamente arraigadas

en los datos de entrenamiento. Su corrección técnica es necesaria pero insuficiente: requiere también un cuestionamiento de la calidad ética de los datos históricos utilizados y, en algunos casos, la prohibición del uso de determinadas categorías de datos para sistemas que afectan a derechos fundamentales.

América Latina se encuentra en una situación de doble vulnerabilidad ante los riesgos de la justicia penal automatizada: las desigualdades estructurales de sus sistemas de justicia hacen que los sesgos algorítmicos sean potencialmente más pronunciados, mientras que la menor capacidad institucional de sus organismos de supervisión hace más difícil el control efectivo. La adopción de marcos regulatorios vinculantes, la creación de organismos de auditoría independiente y la incorporación a los instrumentos interamericanos de derechos humanos de garantías específicas frente a las decisiones algorítmicas son reformas urgentes que los Estados de la región no pueden seguir postergando.

La propuesta de un estatuto ético-jurídico de la IA en la justicia penal, articulado en torno a los ejes de usos prohibidos, usos condicionados, derechos de las personas afectadas, obligaciones de los Estados y mecanismos de rendición de cuentas, representa el horizonte normativo hacia el que debe orientarse la reforma del derecho y la gobernanza tecnológica en este ámbito. Su construcción es una tarea que exige la convergencia de la dogmática jurídico-penal, la ética aplicada, la ciencia de la computación y la participación democrática de los grupos más afectados por las decisiones algorítmicas.

Las líneas de investigación futura que se derivan del presente análisis incluyen: el estudio longitudinal del impacto del Reglamento de IA de la UE sobre la práctica de la justicia penal en los Estados miembros; el análisis de la jurisprudencia interamericana sobre IA y derechos humanos a medida que esta se desarrolle; la evaluación empírica del impacto de los sistemas de IA penal sobre grupos específicos en el contexto latinoamericano mediante análisis de datos de expedientes judiciales; y el desarrollo de propuestas técnicas concretas de auditoría ética algorítmica adaptadas a las condiciones institucionales de la región.

CAPITULO VI

Perspectivas Futuras: Reformas Legislativas y Política Criminal

Perspectivas Futuras: Reformas Legislativas y Política Criminal

YEFERSON CRISTIAN CAICEDO VALENCIA[6]

RESUMEN

El presente capítulo examina las perspectivas futuras de reforma legislativa y de política criminal que deben orientar la adaptación del derecho penal frente a la expansión de la inteligencia artificial (IA). A partir de los fundamentos conceptuales, los problemas de responsabilidad, las tipologías delictivas, la evidencia algorítmica y la dimensión ético-iusfundamental desarrollados en los capítulos precedentes de esta obra, se analizan los escenarios normativos emergentes, los modelos de reforma legislativa comparados y las propuestas de política criminal que la doctrina especializada ha formulado para afrontar los desafíos sistémicos que la IA plantea al ius puniendi. Mediante una metodología de revisión bibliográfica sistemática y análisis comparado de fuentes secundarias, se identifican cuatro ejes de reforma indispensables: la actualización de los tipos penales, la reformulación de los criterios de imputación, la modernización del estatuto procesal de la evidencia algorítmica y la construcción de marcos supranacionales de cooperación. Los resultados evidencian que ningún ordenamiento jurídico ha articulado aún una respuesta sistémica y coherente a estos desafíos, y que la fragmentación normativa global genera riesgos de forum shopping punitivo y de zonas de impunidad algorítmica. Se concluye que la reforma del derecho penal ante la IA requiere un proceso deliberativo inclusivo que preserve los principios garantistas del Estado de Derecho.

Palabras clave: *reformas legislativas; política criminal; inteligencia artificial; derecho penal; tipificación algorítmica; cooperación internacional; imputación penal; gobernanza tecnológica.*

[6] Abogado de los Juzgados y Tribunales de la República del Ecuador, Universidad Guayaquil, Licenciado en Ciencias Sociales y Politicas, Universidad Guayaquil, Especialista en Ciencias penales y Criminalogia, Universidad de Guayaquil, Magister en Derecho Constitucional, Universidad Catolica Santiago de Guayaquil, Magister en Derecho Penal, Universidad Particular de Especialidades Espiritu Santo, Agente Fiscal, Fiscalía General del Estado, correo electrónico: caicedovy@fiscalia.gob.ec.

ABSTRACT

This chapter examines future prospects for legislative reform and criminal policy that should guide the adaptation of criminal law to the expansion of artificial intelligence (AI). Building upon the conceptual foundations, issues of liability, criminal typologies, algorithmic evidence, and the ethical and fundamental rights dimension developed in previous chapters of this work, it analyzes emerging normative scenarios, comparative models of legislative reform, and criminal policy proposals formulated by legal scholars to address the systemic challenges that AI poses to the power to punish (ius puniendi). Through a methodology of systematic literature review and comparative analysis of secondary sources, four essential areas of reform are identified: updating criminal offenses, reformulating the criteria for imputation, modernizing the procedural status of algorithmic evidence, and constructing supranational frameworks for cooperation. The results show that no legal system has yet articulated a systemic and coherent response to these challenges, and that global regulatory fragmentation creates risks of punitive forum shopping and areas of algorithmic impunity. It is concluded that reforming criminal law in the face of AI requires an inclusive deliberative process that preserves the fundamental principles of the rule of law.

Keywords: *legislative reforms; criminal policy; artificial intelligence; criminal law; algorithmic criminalization; international cooperation; criminal prosecution; technological governance.*

INTRODUCCIÓN

Los cinco capítulos precedentes de esta obra han construido un panorama sistemático y progresivo de los desafíos que la inteligencia artificial plantea al derecho penal contemporáneo. El Capítulo I estableció los fundamentos conceptuales que articulan la intersección entre tecnología y ius puniendi, identificando las tensiones estructurales entre eficiencia algorítmica y garantismo penal. El Capítulo II analizó en profundidad la crisis de imputación que los sistemas autónomos generan en las categorías clásicas de la responsabilidad penal, documentando el denominado vacío de responsabilidad que ningún ordenamiento ha logrado colmar de manera satisfactoria. El Capítulo III cartografió las nuevas tipologías delictivas vinculadas a la IA desde el fraude algorítmico hasta los ciberataques autónomos y los crímenes de identidad sintética evidenciando el desfase normativo de primera magnitud entre las conductas posibles y los tipos penales vigentes. El Capítulo IV abordó la problemática de la prueba digital y la evidencia algorítmica en el juicio penal, revelando una brecha normativa entre la sofisticación técnica de los sistemas generadores de evidencia y la capacidad del derecho procesal penal para someterlos a contradicción efectiva. El Capítulo V examinó la dimensión ético-iusfundamental de la justicia penal automatizada, estableciendo que la compatibilidad con el Estado de Derecho exige marcos éticos vinculantes, supervisión humana significativa y rendición de cuentas algorítmica.

El presente capítulo afronta la pregunta que lógicamente corona este itinerario analítico: ¿cómo debería adaptarse el derecho penal frente a la IA? No se trata de una pregunta retórica ni de alcance meramente académico. Cada año que transcurre sin reformas legislativas adecuadas es un año en que sistemas de IA se despliegan en el ámbito de la justicia penal sin marcos normativos que garanticen la protección de los derechos fundamentales, en que conductas delictivas habilitadas por IA quedan impunes por ausencia de tipos aplicables, en que víctimas de crímenes algorítmicos ven frustradas sus expectativas de reparación y en que la fragmentación normativa global genera oportunidades estructurales de forum shopping punitivo para los actores del ecosistema tecnológico más sofisticados (Hildebrandt, 2020; Sieber, 2020; Floridi et al., 2022).

La relevancia de esta pregunta ha sido subrayada desde distintas perspectivas doctrinales. Desde la dogmática penal continental, Roxin (2020) y Mir Puig (2023) han insistido en que la adaptación del derecho penal a las transformaciones tecnológicas no puede hacerse a costa

de los principios garantistas del Estado de Derecho, y que cualquier reforma debe partir del respeto escrupuloso al principio de culpabilidad, al principio de legalidad y al principio de proporcionalidad. Desde el derecho penal comparado, Sieber (2020) y Clough (2021) han documentado la heterogeneidad de las respuestas nacionales y la urgencia de la coordinación internacional. Desde la perspectiva de la regulación tecnológica, Hildebrandt (2020) y Pasquale (2020) han advertido sobre los riesgos de una regulación reactiva y fragmentaria que se limita a adaptaciones cosméticas de las categorías existentes sin abordar los problemas estructurales que la IA plantea. Desde América Latina, Palazzi (2022), Sain (2021) y Ortega & Salinas (2023) han señalado la especial vulnerabilidad de los ordenamientos de la región ante los crímenes algorítmicos y la urgencia de reformas que actualicen marcos normativos concebidos para el cibercrimen de primera y segunda generación.

El capítulo se estructura en las siguientes partes. Tras esta introducción y la sección metodológica, los resultados se organizan en cuatro grandes ejes: el diagnóstico del estado actual de los marcos normativos y sus déficits sistémicos; los modelos de reforma legislativa comparados y sus fortalezas y debilidades; las propuestas de política criminal orientadas a la prevención y respuesta ante los crímenes algorítmicos; y los escenarios de cooperación internacional y armonización normativa. La discusión y las conclusiones sintetizan los hallazgos, identifican las limitaciones del análisis y trazan las líneas de investigación futura. El capítulo cierra la secuencia lógica de la obra articulando, desde una perspectiva de lege ferenda, los principios y las propuestas concretas que deben orientar la adaptación del derecho penal a la era de la inteligencia artificial.

MATERIAL Y MÉTODOS

El presente capítulo adopta un diseño de investigación de naturaleza teórico-documental, coherente con la estrategia metodológica del conjunto de la obra. El diseño combina la revisión bibliográfica sistemática con el método comparativo de fuentes secundarias y el análisis de política legislativa (policy analysis), abordando la literatura científica especializada en derecho penal comparado, regulación de la inteligencia artificial, criminología crítica de la tecnología y gobernanza de riesgos tecnológicos, así como los instrumentos normativos internacionales, regionales y nacionales relevantes, las propuestas legislativas en proceso de tramitación y los documentos de política pública de organismos internacionales de referencia. Esta combinación metodológica es la más adecuada para el objeto de estudio, dado que las

perspectivas futuras de reforma legislativa constituyen un terreno en construcción que requiere tanto el análisis riguroso del estado del arte normativo como la evaluación crítica de las propuestas en discusión.

El método comparativo se emplea en su modalidad funcional (Zweigert & Kötz, 2020), partiendo de los problemas jurídicos que la IA plantea de manera común en distintos ordenamientos para examinar las soluciones que cada sistema jurídico ha articulado o está articulando. El análisis de política legislativa se apoya en el marco analítico de la evaluación de impacto normativo, que examina las propuestas de reforma en función de su capacidad para alcanzar los objetivos de política pública declarados, de su compatibilidad con los principios constitucionales y de derechos humanos aplicables y de su viabilidad técnica e institucional. La integración de ambos métodos permite construir un análisis que articula la dimensión normativo-dogmática con la dimensión empírico-política de las reformas legislativas.

La construcción del corpus documental se realizó mediante búsquedas sistemáticas en las bases de datos Scopus, Web of Science, HeinOnline, Westlaw International, SSRN y EUR-Lex durante el período comprendido entre enero de 2024 y marzo de 2025. Los descriptores de búsqueda incluyeron los términos: "criminal law artificial intelligence reform", "AI criminal legislation comparative", "algorithmic crime legislation reform", "AI criminal policy", "criminal law future AI", "penal reform artificial intelligence", "AI governance criminal law", y sus equivalentes en español, francés y alemán. Se priorizaron artículos de revistas indexadas en cuartiles Q1 y Q2 según el Journal Citation Reports, monografías académicas de editoriales de prestigio, e instrumentos normativos en proceso de tramitación o recientemente aprobados en la Unión Europea, Estados Unidos, el Consejo de Europa y los principales ordenamientos latinoamericanos. El corpus final procesado comprende 215 fuentes documentales, de las cuales 52 fueron seleccionadas como referencias directamente citadas.

El análisis de los materiales seleccionados se llevó a cabo mediante la técnica del análisis documental cualitativo, complementada con la construcción de categorías analíticas propias y la elaboración de cuadros comparativos de marcos normativos. La triangulación de perspectivas disciplinares dogmática jurídico-penal, ciencia política regulatoria, ética tecnológica y criminología, se empleó como estrategia de validación interna del análisis. Entre las principales limitaciones del diseño cabe señalar las siguientes: la acelerada evolución del campo normativo implica que algunos desarrollos legislativos pueden haber progresado o

modificado su contenido después del cierre del corpus; la preponderancia de la literatura en inglés y en el contexto europeo introduce un sesgo geográfico que se ha intentado compensar con la incorporación sistemática de fuentes iberoamericanas; y el carácter prospectivo de parte del análisis implica un nivel de incertidumbre inherente que ha sido explicitado en cada caso mediante el uso de formulaciones condicionadas y la presentación de escenarios alternativos.

RESULTADOS

Diagnóstico del estado normativo global: déficits sistémicos y urgencias de reforma

El desfase estructural entre derecho penal y tecnología algorítmica

El diagnóstico que emerge de la síntesis de los capítulos precedentes y de la revisión bibliográfica específica del presente capítulo es inequívoco: existe un desfase estructural y sistémico entre el derecho penal vigente en la generalidad de los ordenamientos y los desafíos que la inteligencia artificial plantea al sistema de justicia criminal. Este desfase no es el resultado de una deficiencia puntual ni de un retraso momentáneo en la actualización normativa, sino de una discontinuidad cualitativa entre el paradigma tecnológico para el que el derecho penal fue diseñado. La herramienta pasiva al servicio de la voluntad humana y el paradigma emergente, el agente algorítmico con autonomía decisional, opacidad interna y capacidad de aprendizaje adaptativo (Hildebrandt, 2020; Sieber, 2020; Brundage et al., 2021).

La magnitud del desfase puede calibrarse desde tres dimensiones analíticas. La primera es la dimensión de los tipos penales: los tipos vigentes en la mayoría de los ordenamientos fueron concebidos para conductas humanas directas o mediadas por herramientas simples, y no capturan la especificidad técnica y la causalidad difusa de los crímenes algorítmicos. La segunda es la dimensión de la imputación: los criterios de autoría, participación y culpabilidad presuponen un sujeto activo humano con conciencia y voluntad identificables, y no disponen de categorías adecuadas para los supuestos de daño producido por sistemas con autonomía emergente y cadenas causales multi-actor. La tercera es la dimensión procesal: los estándares de admisibilidad probatoria, la cadena de custodia y el principio de contradicción fueron diseñados para la prueba física y testimonial, y resultan insuficientes para la evidencia algorítmica opaca y estadísticamente compleja (Casey, 2019; Kerr, 2020; Dressel & Farid, 2022).

La revisión de los marcos normativos actualmente vigentes permite identificar cuatro categorías de déficits sistémicos. El primer déficit es el de tipicidad: los crímenes algorítmicos de tercera generación (fraude por IA generativa, deepfakes delictivos, ciberataques autónomos, manipulación de sistemas de decisión, desinformación algorítmica masiva) no encuentran cobertura típica satisfactoria en ningún ordenamiento jurídico analizado, y la respuesta estándar de aplicar extensivamente los tipos existentes genera inseguridad jurídica y violaciones del principio de legalidad (Dremliuga & Kucherov, 2022; Chesney & Citron, 2022).

El segundo déficit es el de imputabilidad: el vacío de responsabilidad documentado en los supuestos de daño producido por sistemas de alta autonomía persiste en todos los sistemas jurídicos analizados, sin que ninguno haya articulado una solución que preserve simultáneamente el principio de culpabilidad y la efectividad de la protección a las víctimas (Scherer, 2021; Hallevy, 2022).

El tercer déficit es el de evidencia: la ausencia de estándares normativos específicos para la admisibilidad, validación y contradicción de la evidencia algorítmica expone los procesos penales al riesgo de condenas fundadas en prueba técnicamente no verificada (Raeini et al., 2023; Cheng et al., 2024).

El cuarto déficit es el de cooperación: la transnacionalidad estructural de los crímenes algorítmicos choca con la territorialidad del derecho penal y con la lentitud de los mecanismos de cooperación jurídica internacional, generando zonas de impunidad exploradas sistemáticamente por los actores criminales más sofisticados (Clough, 2021; Europol, 2023).

Déficit sistémico	Manifestación principal	Ordenamientos más afectados	Urgencia
Tipicidad	Crímenes algorítmicos sin tipo penal aplicable; extensión forzada de tipos clásicos	América Latina (todos); España; algunos EEMM UE	Alta

Déficit sistémico	Manifestación principal	Ordenamientos más afectados	Urgencia
	genera inseguridad jurídica		
Imputabilidad	Vacío de responsabilidad en sistemas de alta autonomía; cadena multi-actor sin criterios claros de distribución	Global: ningún sistema ha resuelto plenamente	Muy alta
Evidencia algorítmica	Ausencia de estándares de admisibilidad, validación y contradicción de la prueba IA	América Latina; Reino Unido; países sin regulación específica	Alta
Cooperación internacional	Territorialidad penal vs. transnacionalidad de los crímenes algorítmicos; mecanismos de cooperación lentos	Global: especialmente severo en AL	Alta
Formación institucional	Déficit de capacidades técnicas en	América Latina; países en desarrollo	Media-alta

Déficit sistémico	Manifestación principal	Ordenamientos más afectados	Urgencia
	fiscalías, judicaturas y cuerpos policiales		

Elaboración propia a partir de Sieber (2020); Clough (2021); Dremliuga & Kucherov (2022); Europol (2023); Palazzi (2022); Scherer (2021).

El fenómeno del forum shopping punitivo algorítmico

Una de las consecuencias más preocupantes del desfase normativo global es la emergencia del fenómeno que la doctrina ha comenzado a denominar forum shopping punitivo algorítmico: la estrategia mediante la cual los actores del ecosistema tecnológico (desarrolladores, desplegadores y, en su caso, actores criminales) ubican sus actividades en las jurisdicciones con marcos normativos más laxos, aprovechando la fragmentación regulatoria global para minimizar su exposición al derecho penal. Este fenómeno es análogo al que se produce en el ámbito del derecho fiscal internacional con los paraísos fiscales, pero con implicaciones específicas para la seguridad jurídica y la protección de los derechos fundamentales (Sieber, 2020; Europol, 2023).

La literatura identifica tres modalidades del forum shopping punitivo algorítmico. La primera es la deslocalización de servidores e infraestructura: los actores que desarrollan o despliegan sistemas de IA con finalidades delictivas o en zonas grises de legalidad ubican su infraestructura técnica en jurisdicciones donde no existe tipificación penal de esas conductas o donde la capacidad de persecución es mínima. La segunda es la estructura corporativa multijurisdiccional: los grupos corporativos que desarrollan sistemas de IA de alto riesgo distribuyen sus operaciones entre múltiples jurisdicciones de manera que ninguna de ellas pueda ejercer competencia penal plena sobre el conjunto de la cadena de responsabilidad. La tercera es el aprovechamiento de las asimetrías en los mecanismos de cooperación internacional: los actores criminales sofisticados calculan los tiempos de la asistencia judicial mutua y diseñan sus operaciones para que los datos y la evidencia relevante hayan

desaparecido o devenido inaccesibles antes de que los mecanismos de cooperación puedan activarse (Clough, 2021; Consejo de Europa, 2022; Europol, 2023).

Modelos de reforma legislativa: análisis comparado

El modelo de reforma sistémica: la apuesta europea

El Reglamento de Inteligencia Artificial de la Unión Europea (2024/1689) representa el ejemplo más avanzado a nivel global de reforma normativa sistémica en el ámbito de la IA. Aunque su ámbito material es primordialmente preventivo y regulatorio, no penal en sentido estricto, sus disposiciones tienen implicaciones penales indirectas de primera magnitud y configuran el marco de referencia más comprehensivo disponible para orientar las reformas legislativas en el ámbito penal (Comisión Europea, 2024; Floridi et al., 2022).

Las características definitorias del modelo europeo de reforma son cuatro. La primera es el enfoque basado en riesgos: en lugar de regular tecnologías específicas o conductas concretas, el Reglamento de IA establece categorías de riesgo (inaceptable, alto, limitado, mínimo) y asocia a cada categoría un conjunto de requisitos proporcionales al riesgo que el sistema presenta para los derechos fundamentales. Esta arquitectura regulatoria tiene la ventaja de la flexibilidad tecnológica: no queda obsoleta ante la aparición de nuevas tecnologías, sino que se aplica a cualquier sistema que presente el nivel de riesgo definido, con independencia de la tecnología subyacente. La segunda característica es la obligatoriedad jurídica: a diferencia de la mayoría de los marcos normativos existentes sobre IA, el Reglamento europeo es un instrumento de derecho primario de aplicación directa y obligatoria en los 27 Estados miembros, con sanciones de hasta 35 millones de euros o el 7% del volumen de negocio global para las infracciones más graves. La tercera característica es la integración en el ecosistema normativo existente: el Reglamento de IA no sustituye sino que se superpone y complementa al Reglamento General de Protección de Datos (RGPD), a la Directiva NIS2, al Reglamento de Servicios Digitales (DSA) y a los marcos sectoriales de responsabilidad civil. La cuarta característica es la orientación hacia la supervisión humana: el Reglamento establece con carácter general la exigencia de supervisión humana significativa para los sistemas de alto riesgo, lo que en el ámbito penal significa que ninguna decisión relevante para los derechos de las personas puede adoptarse exclusivamente sobre la base de conclusiones algorítmicas.

Las propuestas doctrinales para completar el modelo europeo en la dimensión penal convergen en torno a tres líneas de reforma. La primera es la tipificación penal a nivel de directiva de las conductas que vulneran las prohibiciones absolutas del Reglamento de IA (sistemas de puntuación social, manipulación subliminal, identificación biométrica no autorizada), estableciendo un estándar mínimo de sanción penal armonizado para todos los Estados miembros. La segunda es la armonización de los criterios de responsabilidad penal corporativa aplicables a las empresas del ecosistema de la IA que incurran en defectos de organización que contribuyan a la producción de daños penalmente relevantes. La tercera es la creación de un estatuto procesal penal específico para la evidencia algorítmica, que establezca estándares de admisibilidad, validación y contradicción aplicables de manera uniforme en los procesos penales de todos los Estados miembros (Sieber, 2020; Gómez-Jara Díez, 2021; Nieto Martín, 2021).

El modelo de reforma incremental: la adaptación angloamericana

El modelo angloamericano de reforma legislativa ante la IA se caracteriza por la adaptación incremental de los marcos normativos existentes mediante la legislación específica por modalidades, el desarrollo jurisprudencial creativo y la autorregulación del sector. Este modelo refleja la tradición del common law de evolución normativa a través de la casuística judicial y la adaptación contextualizada de los principios generales a los nuevos supuestos de hecho, antes que mediante la codificación sistemática de nuevas categorías dogmáticas (Scherer, 2021; Vladeck, 2022).

En los Estados Unidos, la adaptación incremental se ha manifestado en una pluralidad de iniciativas legislativas de alcance específico: el DEFIANCE Act (2024) criminalizó la producción y distribución de deepfakes sexuales no consentidos; el National Cybersecurity Strategy (2023) redefinió las prioridades de persecución penal en el ámbito del cibercrimen potenciado por IA; varias iniciativas estatales han regulado el uso del reconocimiento facial por las fuerzas del orden y los deepfakes electorales; y la jurisprudencia federal ha comenzado a aplicar los tipos de fraude, manipulación de mercados y obstaculización a la justicia a los primeros supuestos de crímenes algorítmicos documentados ante los tribunales. La virtud del modelo incremental es su adaptabilidad y su pragmatismo. Su debilidad es la incoherencia sistémica que genera a largo plazo: la acumulación de reformas sectoriales sin una arquitectura normativa unificadora produce marcos jurídicos fragmentados, con lagunas y superposiciones que generan

inseguridad jurídica para todos los actores del ecosistema. Como ha señalado Hildebrandt (2020), la regulación por defecto tecnológico, en la que son las decisiones de diseño de los ingenieros las que determinan de facto el contenido de la norma, es el resultado más probable de una estrategia de reforma puramente incremental sin un marco sistémico que la oriente.

El modelo latinoamericano: entre la modernización urgente y la dependencia normativa

La situación de los ordenamientos latinoamericanos en el proceso de adaptación legislativa del derecho penal ante la IA puede ser caracterizada como de modernización urgente condicionada por la dependencia normativa. La urgencia deriva de la brecha ya documentada entre los marcos normativos vigentes diseñados para el cibercrimen de primera y segunda generación, y los crímenes algorítmicos de tercera generación que las redes criminales ya despliegan en la región. La dependencia normativa deriva de la tendencia histórica de los ordenamientos latinoamericanos a importar modelos normativos europeos o norteamericanos sin los necesarios procesos de adaptación a las especificidades de los contextos locales (Palazzi, 2022; Sain, 2021; Ortega & Salinas, 2023).

El análisis comparado de los procesos legislativos en curso en la región identifica cuatro tendencias dominantes. La primera es la incorporación progresiva al Convenio de Budapest sobre Ciberdelincuencia: Brasil, Chile, Colombia y Argentina ya son partes o están en proceso avanzado de adhesión, lo que implica la actualización de sus marcos de cooperación internacional y la armonización con los estándares mínimos del Convenio. La segunda tendencia es la elaboración de políticas nacionales de IA que incluyen referencias a la seguridad y la justicia. La tercera tendencia es la reforma específica de los Códigos Penales en materia de delitos informáticos: Chile aprobó la Ley 21.459 (2022), que modernizó sustantivamente su marco de ciberdelitos; Brasil ha tramitado reformas a la Ley 14.155/2021; y Colombia ha actualizado algunos aspectos de la Ley 1273/2009. La cuarta tendencia, la más relevante para el presente análisis, es la elaboración de proyectos de ley específicos de regulación de la IA que en varios países incluyen disposiciones sobre responsabilidad y seguridad (Palazzi, 2022; Ortega & Salinas, 2023).

Sin embargo, el análisis más riguroso de estos procesos legislativos revela que, en su estado actual, son insuficientes para afrontar los desafíos que los crímenes algorítmicos plantean. Los proyectos de ley de IA en tramitación en la mayoría de los países de la región tienen un

enfoque predominantemente administrativo y de política pública, sin disposiciones sustantivas sobre la responsabilidad penal de los actores del ecosistema de la IA. Y la incorporación al Convenio de Budapest, aunque esencial, solo actualiza el marco de cooperación internacional sin resolver los problemas sustantivos de tipificación y de imputación (Sain, 2021; Ortega & Salinas, 2023).

País	Estado reforma penal IA	Convenio Budapest	Marco regulatorio IA	Evaluación global
Brasil	L. 14.155/2021; debate L. IA en Congreso	Proceso de adhesión en curso	Marco de IA en elaboración (2024)	En transición; déficits en tipificación algorítmica
Chile	L. 21.459 (2022) - más actualizada de la región	Ratificado (2023)	Política Nacional IA 2021	Avanzado; aun sin tipos específicos para IA generativa
Colombia	L. 1273/2009; jurisprudencia CSJ; proyectos en curso	Ratificado (2021)	Política IA 2022; proyectos ley IA	Intermedio; framework sólido pero desactualizado
Argentina	LL. 26.388 y 26.904; reformas parciales	Proceso adhesión en curso	Marco regulatorio IA en debate	Fragmentado; múltiples proyectos sin consenso
Ecuador	COIP arts. 229-234; sin	Ratificado (2024)	Borrador política IA 2023	Rezagado; actualización urgente

País	Estado reforma penal IA	Convenio Budapest	Marco regulatorio IA	Evaluación global
	reforma específica IA			
México	CPF arts. 211-211b; sin reformas recientes	Sin adhesión	Iniciativas legislativas múltiples, sin aprobación	Muy rezagado; fragmentación normativa grave

Elaboración propia a partir de Palazzi (2022); Sain (2021); Ortega & Salinas (2023); Clough (2021); Consejo de Europa (2022).

Propuestas de reforma del derecho penal sustantivo

Actualización de los tipos penales: hacia una tipología de crímenes algorítmicos

El eje más urgente de la reforma legislativa en el ámbito del derecho penal sustantivo es la actualización de los tipos penales para dar cobertura adecuada a los crímenes algorítmicos de tercera generación. La doctrina especializada ha convergido en torno a dos grandes estrategias legislativas que no son mutuamente excluyentes: la estrategia de la agravación (introducir circunstancias agravantes específicas vinculadas al uso de IA en los tipos penales existentes) y la estrategia de la tipificación específica (crear tipos penales autónomos para las conductas que son cualitativamente distintas de las ya tipificadas) (Dremliuga & Kucherov, 2022; Sieber, 2020; Chesney & Citron, 2022).

La estrategia de la agravación presenta la ventaja de la compatibilidad con los marcos normativos existentes y la menor exigencia de reforma legislativa. Su implementación puede realizarse mediante la introducción en los tipos base de estafa, fraude, daños informáticos, falsedad y delitos contra la intimidad de una circunstancia agravante que incremente la penalidad cuando la conducta se cometa mediante el uso de sistemas de IA. Este modelo permite capturar el mayor desvalor de la conducta algorítmica (su escala, su sofisticación

técnica, su capacidad de adaptación y la dificultad de su detección y atribución) sin necesidad de crear categorías dogmáticas radicalmente nuevas.

La estrategia de la tipificación específica es metodológicamente más exigente pero normativamente más precisa para las conductas que presentan una especificidad técnica y jurídica que no puede capturarse adecuadamente mediante la agravación de los tipos existentes. La doctrina identifica al menos cinco categorías de conductas que requieren tipificación autónoma por su especificidad. La primera es la manipulación o envenenamiento de datos de entrenamiento de sistemas de IA con finalidad delictiva (data poisoning), que no tiene equivalente en la tipificación clásica de los daños informáticos y que puede afectar de manera sistémica y silenciosa a múltiples bienes jurídicos protegidos. La segunda es la creación y distribución de agentes de IA autónomos con finalidad delictiva. La tercera es la explotación delictiva de los sistemas de reconocimiento biométrico mediante ataques adversariales o suplantación biométrica de alta sofisticación. La cuarta es la desinformación algorítmica orquestada (producción industrial de contenido sintético falso mediante IA para interferir en procesos electorales o desestabilizar instituciones democráticas). La quinta es la gestión irresponsable de sistemas de IA de alto riesgo que produce daños individuales o colectivos graves (Chesney & Citron, 2022; Dremliuga & Kucherov, 2022; Ferrara, 2023).

Reformulación de los criterios de imputación penal

La segunda dimensión de la reforma del derecho penal sustantivo es la reformulación de los criterios de imputación penal para los supuestos de daño producido por sistemas autónomos de IA. Como se analizó en el Capítulo II, el vacío de responsabilidad que estos supuestos generan no puede resolverse mediante la simple aplicación extensiva de las categorías existentes de autoría y participación, y requiere el desarrollo de nuevas categorías dogmáticas o la adaptación sustantiva de las existentes. La propuesta que ha alcanzado mayor consenso en la doctrina comparada es la de la imputación diferenciada por niveles de autonomía del sistema, combinada con la clarificación legislativa de los deberes de diligencia exigibles a cada actor de la cadena (Scherer, 2021; Sieber, 2020; Gómez-Jara Díez, 2021).

El modelo de imputación diferenciada por niveles de autonomía distingue tres escalones en la cadena de responsabilidad penal según el grado de autonomía del sistema de IA involucrado. En el primer escalón, sistemas de baja autonomía operados directamente por el

usuario, la responsabilidad penal del usuario se establece mediante las categorías ordinarias de autoría directa o instrumental, sin necesidad de modificación dogmática. En el segundo escalón, sistemas de autonomía media que actúan con independencia operacional dentro de parámetros fijados por el desplegador, la responsabilidad penal del desplegador se articula mediante la figura del garante digital (Sieber, 2020), quien tiene la obligación de supervisar el funcionamiento del sistema y de intervenir cuando sus decisiones superen los parámetros de riesgo aceptables; el incumplimiento de esta obligación configura el tipo omisivo de comisión por omisión. En el tercer escalón, sistemas de alta autonomía con comportamiento emergente no predecible, la responsabilidad penal individual resulta problemática por ausencia del elemento subjetivo imprudente respecto al resultado específico, y debe complementarse con un régimen de responsabilidad cuasi-objetiva del fabricante o de un fondo de compensación sectorial que garantice la reparación a las víctimas sin sacrificar el principio de culpabilidad.

La concreción legislativa de este modelo requiere al menos tres reformas normativas específicas. La primera es la codificación en los Códigos Penales o en leyes especiales de los deberes de diligencia exigibles al desarrollador, al fabricante, al desplegador y al usuario de sistemas de IA de alto riesgo, con referencia explícita a los estándares técnicos certificados por organismos de normalización competentes. Esta codificación transforma la pregunta dogmática de qué nivel de diligencia era exigible en una pregunta jurídica con respuesta normativa predefinida, lo que incrementa la previsibilidad y la seguridad jurídica. La segunda reforma es la introducción explícita en los códigos penales de la figura del defecto de organización en el desarrollo y despliegue de sistemas de IA de alto riesgo como fundamento de la responsabilidad penal corporativa. La tercera reforma es la clarificación de los criterios de atribución del elemento subjetivo imprudente en los supuestos de IA autónoma, estableciendo presunciones iuris tantum de negligencia respecto al fabricante o desplegador cuando el daño producido por el sistema esté dentro del ámbito de riesgo que los estándares técnicos aplicables pretendían prevenir (Scherer, 2021; Vladeck, 2022; Gómez-Jara Díez, 2021).

El principio de legalidad ante el desafío de la evolución tecnológica

Uno de los principales obstáculos para la reforma del derecho penal sustantivo en materia de IA es la tensión entre el principio de legalidad que exige que los tipos penales sean suficientemente precisos para que sus destinatarios puedan prever las consecuencias jurídicas

de sus conductas, y la aceleración tecnológica, que hace que los tipos altamente técnicos queden obsoletos con mayor rapidez de la que los procesos legislativos pueden actualizar. Esta tensión ha sido señalada como uno de los problemas centrales de la reforma por autores como Sieber (2020) y Hildebrandt (2020), y su resolución requiere un enfoque legislativo específicamente adaptado a la naturaleza del objeto regulado.

La doctrina comparada ha identificado tres técnicas legislativas que pueden conciliar la precisión que el principio de legalidad exige con la flexibilidad que la dinámica tecnológica impone. La primera es la técnica de las cláusulas abiertas con remisión a estándares técnicos: los tipos penales se redactan mediante referencias a comportamientos que infringen estándares técnicos de seguridad o de diligencia establecidos por organismos de normalización competentes (como el NIST estadounidense, el CEN europeo o las normas ISO/IEC), cuyo contenido puede actualizarse mediante procedimientos más ágiles que los legislativos sin menoscabar la certeza normativa. Esta técnica, utilizada con éxito en el derecho penal del medioambiente y en el derecho penal económico, es la más prometedora para el ámbito de la IA. La segunda técnica es la de los tipos de peligro abstracto basados en categorías funcionales: en lugar de describir las conductas típicas mediante referencias a tecnologías específicas que pueden quedar obsoletas, los tipos se redactan en términos funcionales que capturan el peligro que la conducta representa independientemente de la tecnología utilizada. La tercera técnica es la de las cláusulas de actualización normativa periódica: los propios textos legislativos establecen mecanismos de revisión y actualización periódica, por ejemplo, cada tres o cinco años, que garantizan su adecuación continua al estado del arte tecnológico sin necesidad de tramitaciones legislativas ordinarias para cada modificación (Sieber, 2020; Dremliuga & Kucherov, 2022).

Ámbito de reforma	Medida propuesta	Referencia legislativa comparada	Compatibilidad con garantías
Tipificación de crímenes algorítmicos	Circunstancias agravantes por uso de IA + tipos autónomos para conductas	DEFIANCE Act (EE.UU., 2024); L. 21.459 Chile (2022); Propuesta Directiva UE tipificación IA	Alta si se respeta taxatividad; aplicar técnicas

Ámbito de reforma	Medida propuesta	Referencia legislativa comparada	Compatibilidad con garantías
	específicas (data poisoning, desinformación orquestada, agentes IA maliciosos)		de cláusulas abiertas con remisión a estándares
Criterios de imputación	Codificación de deberes de diligencia por nivel de autonomía; defecto de organización IA en responsabilidad corporativa; presunciones iuris tantum en alta autonomía	Reglamento UE IA arts. 9-17; GDPR art. 22; modelos de compliance penal corporativo (España, Francia)	Media-alta; requiere salvar principio de culpabilidad mediante culpa in vigilando o in eligendo
Principio de legalidad y tecnicidad	Remisión a estándares técnicos certificados (NIST, ISO/IEC, CEN); tipos de peligro abstracto funcionales; mecanismos de revisión periódica	Derecho penal ambiental europeo; NIS2 art. 21; modelo NIST AI RMF	Alta si los estándares son accesibles, públicos y previsibles
Responsabilidad corporativa penal	Extensión del defecto de organización IA; obligaciones de compliance algorítmico;	CPE art. 31 bis; Francia: L. Sapin II; UK Bribery Act; propuestas OCDE gobernanza IA	Alta si se mantienen garantías procesales corporativas

Ámbito de reforma	Medida propuesta	Referencia legislativa comparada	Compatibilidad con garantías
	sanciones acumulables (penal + administrativa)		

Elaboración propia a partir de Sieber (2020); Dremliuga & Kucherov (2022); Scherer (2021); Vladeck (2022); Gómez-Jara Díez (2021); Comisión Europea (2024).

Reformas del derecho procesal penal: el estatuto de la evidencia algorítmica

Hacia un estatuto probatorio específico para la evidencia algorítmica

La reforma del derecho procesal penal en materia de evidencia algorítmica es tan urgente como la reforma del derecho sustantivo, aunque ha recibido menor atención en los procesos legislativos en curso. La ausencia de un estatuto probatorio específico para la evidencia algorítmica expone los procesos penales al doble riesgo de la exclusión acrítica de prueba técnicamente valiosa (por aplicación de estándares de admisibilidad diseñados para la prueba física) y de la admisión acrítica de prueba técnicamente deficiente, por la ilusión de objetividad que la aparente precisión matemática de los sistemas algorítmicos proyecta sobre los juzgadores, (Raeini et al., 2023; Cheng et al., 2024; Dressel & Farid, 2022).

Las propuestas doctrinales para la construcción de un estatuto probatorio específico para la evidencia algorítmica convergen en torno a un conjunto de principios normativos. El primer principio es el de validación técnica previa a la admisión: todo sistema de IA cuyas conclusiones se pretendan utilizar como evidencia en un proceso penal debe haber sido previamente validado mediante evaluación técnica independiente que acredite su fiabilidad, su tasa de error en condiciones similares al caso y la ausencia de sesgos discriminatorios. Esta validación debe realizarse con carácter previo a la admisión de la evidencia y no puede ser suplida por la declaración del fabricante ni por el mero hecho de que el sistema sea de uso policial ordinario. El segundo principio es el de auditabilidad: la defensa tiene derecho a que peritos de su designación puedan examinar el sistema que generó la evidencia, replicar las condiciones de su generación y someter sus conclusiones a prueba de contraste. El tercer principio es el de explicabilidad local: para que la conclusión de un sistema de IA sea admisible como prueba de cargo determinante, el sistema debe ser capaz de proporcionar una

explicación comprensible de los factores que determinaron esa conclusión en el caso concreto. Cuando el sistema no pueda proporcionar esa explicación, la conclusión algorítmica podrá utilizarse únicamente como elemento auxiliar de la valoración probatoria, nunca como prueba de cargo autónoma. El cuarto principio es el de igualdad de armas técnica: los sistemas de justicia penal deben garantizar que los servicios de defensa pública dispongan de acceso a peritos técnicos especializados en sistemas de IA y de los recursos necesarios para ejercer efectivamente el derecho de contradicción ante la evidencia algorítmica (Casey, 2019; Raeini et al., 2023; Cheng et al., 2024; Ferguson, 2021).

La cadena de custodia en la era algorítmica: nuevos protocolos

La modernización de los protocolos de cadena de custodia para la evidencia algorítmica es otra dimensión indispensable de la reforma procesal penal. Los protocolos de cadena de custodia tradicionales, basados en la documentación de los accesos físicos a soportes de almacenamiento y en la verificación de integridad mediante funciones hash, son condición necesaria pero no suficiente para garantizar la integridad de la evidencia algorítmica. A ellos deben añadirse nuevos elementos específicos: la documentación del entorno técnico en el momento de la generación de la evidencia (versión del software, parámetros de configuración, condiciones de los datos de entrada); el registro de todas las interacciones entre el sistema de IA y los datos durante el proceso de análisis; la preservación de los logs de funcionamiento del sistema que permitan la reconstrucción del proceso de decisión; y la documentación del histórico de actualizaciones del sistema que puedan haber modificado su comportamiento entre el momento del análisis y el momento del juicio (Karie & Venter, 2020; Maras & Alexandrou, 2023).

La implementación de estos nuevos estándares de cadena de custodia requiere reformas en tres planos. En el plano normativo, los Códigos de Procedimiento Penal deben actualizar sus disposiciones sobre evidencia digital para incluir los requisitos específicos de la evidencia algorítmica, estableciendo las consecuencias procesales del incumplimiento (nulidad de la evidencia, reducción de su valor probatorio, inversión de la carga de la prueba sobre su integridad). En el plano institucional, los organismos forenses de los cuerpos de seguridad y de las fiscalías deben desarrollar protocolos internos de recogida, preservación y análisis de evidencia algorítmica. En el plano técnico, los sistemas de IA utilizados por las fuerzas del orden deben ser diseñados con capacidades de logging y de auditoría que faciliten el

cumplimiento de los requisitos de cadena de custodia, lo que implica incluir este requisito entre las especificaciones de los contratos de adquisición de tecnología policial (Casey, 2019; Karie & Venter, 2020).

Política criminal ante la IA: prevención, investigación y sanción

Estrategias de prevención: educación, certificación y diseño seguro

La política criminal ante la IA no puede limitarse a la dimensión reactiva de la persecución penal de los crímenes ya cometidos, sino que debe incorporar una dimensión preventiva robusta articulada en torno a tres líneas de acción. La primera línea es la educación y formación: la prevención de los crímenes algorítmicos requiere que los ciudadanos, las empresas y las instituciones públicas dispongan de la alfabetización tecnológica necesaria para identificar y reportar conductas potencialmente delictivas vinculadas a la IA, y que los profesionales del derecho, la seguridad y la justicia dispongan de las competencias técnicas necesarias para investigarlas y enjuiciarlas. Esta formación debe incorporarse a los planes de estudio de las facultades de Derecho, a los programas de formación continua de magistrados y fiscales y a los programas de capacitación de los cuerpos policiales (Europol, 2023; Ortega & Salinas, 2023).

La segunda línea de prevención es la certificación y auditoría previa al despliegue: los sistemas de IA de alto riesgo utilizados en el ámbito de la seguridad pública, la justicia penal y las infraestructuras críticas deben someterse a procesos de certificación técnica independiente antes de su despliegue, que acrediten su fiabilidad, seguridad y ausencia de sesgos discriminatorios. En los países que no cuentan con el marco normativo europeo como referencia obligatoria, la adopción voluntaria de los estándares de certificación establecidos por el NIST AI Risk Management Framework o por los marcos técnicos de la ISO/IEC 42001 puede constituir un primer paso hacia la estandarización global (NIST, 2023; Comisión Europea, 2024).

La tercera línea de prevención es el diseño seguro por defecto (security by design y privacy by design): la incorporación de los requisitos de seguridad, privacidad y ausencia de sesgos discriminatorios en las fases tempranas del diseño de los sistemas de IA, antes de su despliegue, es significativamente más eficaz y menos costosa que su corrección a posteriori. La política criminal puede incentivar el diseño seguro por defecto mediante tres mecanismos:

la creación de ventajas fiscales o de acceso a contratos públicos para las empresas que acrediten el cumplimiento de estándares de diseño seguro; la imposición de responsabilidad civil objetiva por los daños derivados del incumplimiento de estándares de diseño; y la tipificación penal del lanzamiento al mercado de sistemas de IA de alto riesgo sin haber realizado la evaluación de impacto sobre los derechos fundamentales exigida por los marcos normativos aplicables (Hildebrandt, 2020; Floridi et al., 2022; Comisión Europea, 2024).

Capacidades institucionales para la investigación de crímenes algorítmicos

El segundo pilar de la política criminal ante la IA es el fortalecimiento de las capacidades institucionales de investigación de los crímenes algorítmicos. La persecución penal efectiva de los crímenes de tercera generación requiere competencias técnicas especializadas en las unidades de cibercrimen de los cuerpos policiales, en las fiscalías especializadas y en los laboratorios forenses, así como recursos tecnológicos actualizados que permitan el análisis de sistemas de IA complejos. En la mayoría de los países, el déficit de estas capacidades es uno de los principales factores que explican la impunidad de los crímenes algorítmicos, con independencia del estado de los marcos normativos aplicables (Europol, 2023; Ferguson, 2021).

La creación de unidades especializadas en crímenes algorítmicos dentro de los cuerpos policiales y las fiscalías es una prioridad de política criminal de primer orden que la doctrina comparada recomienda de manera prácticamente unánime. Estas unidades deben disponer de equipos multidisciplinares que integren juristas especializados en derecho penal de la tecnología, ingenieros expertos en sistemas de IA, analistas forenses digitales y especialistas en cooperación internacional. La Europol, a través de su European Cybercrime Centre (EC3), y el FBI, a través de su Cyber Division, constituyen modelos de referencia cuya replicación —adaptada a las capacidades presupuestarias de cada país— es un objetivo de política criminal que los gobiernos de América Latina deberían incorporar a sus planes de seguridad a mediano plazo (Europol, 2023; Department of Justice, 2023).

La política criminal de sanciones: proporcionalidad, disuasión y rehabilitación

El tercer pilar de la política criminal ante la IA es el diseño de un sistema de sanciones que sea proporcionado a la gravedad de los crímenes algorítmicos, eficazmente disuasorio respecto a los potenciales infractores y compatible con los principios del derecho penal liberal.

La política de sanciones ante los crímenes algorítmicos presenta desafíos específicos: los crímenes algorítmicos pueden tener un impacto masivo en términos del número de víctimas y del daño global causado, mientras que la determinación de la pena individual debe seguir los criterios de proporcionalidad individual que el principio de culpabilidad impone. Además, los actores corporativos del ecosistema tecnológico son, en muchos supuestos, los responsables más relevantes desde la perspectiva de la prevención general, y las penas privativas de libertad diseñadas para infractores individuales son inadecuadas para producir los efectos preventivos deseados sobre entidades corporativas (Gómez-Jara Díez, 2021; Nieto Martín, 2021; Scherer, 2021).

Las propuestas doctrinales de política de sanciones convergen en torno a un modelo combinado que articula tres niveles de respuesta. El primer nivel comprende las sanciones penales individuales: penas privativas de libertad y penas pecuniarias personale, para los supuestos en que sea posible identificar un responsable humano con culpabilidad penal suficiente. El segundo nivel comprende las sanciones penales corporativas: multas de elevada cuantía, inhabilitaciones para contratar con el sector público, suspensión o prohibición de actividades, obligaciones de reparación, para los supuestos de responsabilidad corporativa por defecto de organización. El tercer nivel comprende las sanciones cuasipenales o administrativas de alta intensidad: complementarias a las penales o subsidiarias en los supuestos en que la persecución penal no es viable, para los supuestos de incumplimiento de los requisitos de seguridad, transparencia y no discriminación. La articulación coherente de estos tres niveles requiere una reforma de los mecanismos de cooperación entre el derecho penal y el derecho administrativo sancionador que muy pocos ordenamientos han abordado de manera explícita (Gómez-Jara Díez, 2021; Nieto Martín, 2021).

Cooperación internacional y armonización normativa

El Convenio de Budapest y sus limitaciones ante los crímenes algorítmicos

El Convenio de Budapest sobre Ciberdelincuencia (2001) es el instrumento internacional más ampliamente ratificado en materia de delitos informáticos y el referente normativo fundamental para la cooperación internacional en la persecución de los crímenes del ciberespacio. Su Segundo Protocolo Adicional (2022) ha modernizado los mecanismos de obtención transfronteriza de evidencia digital y ha ampliado los marcos de cooperación entre

fuerzas policiales y entre autoridades judiciales. Sin embargo, tanto el Convenio original como su Protocolo Adicional presentan limitaciones significativas para afrontar los crímenes algorítmicos de tercera generación que hace necesario complementar o actualizar este marco (Consejo de Europa, 2022; Clough, 2021).

Las principales limitaciones del sistema del Convenio de Budapest ante los crímenes algorítmicos son cuatro. La primera es la desactualización tipológica: el Convenio tipifica los delitos informáticos en categorías concebidas para el paradigma del cibercrimen de primera generación y no captura las especificidades de los crímenes algorítmicos de tercera generación. La segunda limitación es la insuficiencia de los mecanismos de atribución: el Protocolo Adicional ha facilitado el acceso transfronterizo a ciertos datos digitales, pero no ha resuelto el problema de la atribución de los ciberataques autónomos a actores específicos, que sigue siendo el principal cuello de botella de la persecución penal internacional. La tercera limitación es la cobertura geográfica incompleta: varios Estados de relevancia tecnológica mayor, especialmente China y Rusia, no son parte, generando zonas de impunidad geográfica. La cuarta limitación es la velocidad de actualización: el proceso de negociación y ratificación de los instrumentos del Consejo de Europa es intrínsecamente lento en relación con la velocidad de evolución de las tecnologías y las modalidades delictivas (Consejo de Europa, 2022; Clough, 2021; Europol, 2023).

La Convención de las Naciones Unidas sobre Ciberdelincuencia y sus perspectivas

La adopción en 2024 de la Convención de las Naciones Unidas sobre Ciberdelincuencia representa el desarrollo más significativo en el ámbito de la cooperación internacional en materia de cibercrimen desde la adopción del Convenio de Budapest. La Convención de la ONU tiene un potencial de cobertura geográfica significativamente mayor que el Convenio de Budapest, al estar abierta a la adhesión de todos los 193 Estados miembros de las Naciones Unidas. Sin embargo, su texto final ha generado controversias doctrinales importantes sobre su compatibilidad con los estándares de derechos humanos, dado que algunas de sus disposiciones sobre cooperación en la obtención de evidencia digital y sobre la definición de los delitos informáticos podrían ser interpretadas de manera que faciliten la persecución de periodistas, activistas y disidentes políticos en regímenes autoritarios (Europol, 2023; Consejo de Europa, 2022).

La evaluación de las perspectivas de la Convención de la ONU para la persecución de los crímenes algorítmicos es necesariamente matizada. En el aspecto positivo, la Convención amplía el marco de cooperación internacional a un número de Estados significativamente mayor que el Convenio de Budapest, lo que puede reducir las zonas de impunidad geográfica. En el aspecto negativo, su texto no aborda de manera específica los crímenes algorítmicos de tercera generación, replicando en gran medida la estructura tipológica del Convenio de Budapest sin actualización para las conductas vinculadas a la IA; sus disposiciones sobre salvaguardas de derechos humanos son, en opinión de la doctrina mayoritaria, insuficientemente robustas; y el proceso de ratificación e implementación puede tardar años, durante los cuales el desfase normativo continuará (Europol, 2023).

Propuestas de armonización normativa a nivel regional y global

Ante las limitaciones de los instrumentos internacionales existentes, la doctrina comparada ha formulado propuestas de armonización normativa a nivel regional y global que complementen o refuercen el sistema del Convenio de Budapest. En el ámbito regional latinoamericano, Palazzi (2022) y Sain (2021) han propugnado la creación de una Red Iberoamericana de Derecho Penal e Inteligencia Artificial que facilite la coordinación legislativa, el intercambio de mejores prácticas y la capacitación técnica de las instituciones del sistema de justicia. Esta red podría operar en el marco de la Conferencia de Ministros de Justicia de los Países Iberoamericanos (COMJIB), aprovechando los lazos históricos, culturales y jurídicos entre España, Portugal y los países de América Latina para construir un espacio normativo de referencia.

En el ámbito global, la propuesta más influyente es la de Sieber (2020) de elaborar un Protocolo Internacional sobre Crímenes Algorítmicos en el marco del Convenio de Budapest, que actualice las tipologías delictivas para incluir los crímenes de tercera generación y establezca estándares de cooperación específicamente adaptados a la naturaleza de estos crímenes. Este Protocolo podría adoptar un mecanismo de actualización más ágil mediante un órgano técnico permanente de supervisión con capacidad de emitir directrices interpretativas vinculantes sin necesidad de procedimiento legislativo formal en cada Estado parte. La adopción de una convención especializada en crímenes de IA en el marco de la OCDE, cuya experiencia en la elaboración de estándares de gobernanza de la IA constituye un activo

relevante, es otra posibilidad que la doctrina ha mencionado (Sieber, 2020; OCDE, 2023; Europol, 2023).

Instrumento / iniciativa	Ámbito geográfico	Estado actual	Fortalezas	Limitaciones
Convenio de Budapest (2001) + 2.° Protocolo (2022)	68 Estados parte	En vigor; 2.° Prot. en ratificación	Amplia cobertura; marco cooperación consolidado	Tipologías desactualizadas; sin disposiciones IA específicas
Convención ONU Ciberdelincuencia (2024	Universal (potencial)	Adoptada; en proceso ratificación	Potencial cobertura universal	Serias preocupaciones DDHH; sin tipologías algorítmicas
Reglamento UE IA (2024/1689	27 EEMM + EEE	En vigor; aplicación progresiva 2025-2027	Vinculante; sistémico; sanciones elevadas	Sin tipos penales específicos; solo derecho administrativo
Propuesta Red Iberoamericana DPenal-IA	España + Portugal + AL (22 países)	Propuesta doctrinal; sin implementación	Afinidad jurídica y cultural; potencial capacitación	Sin base institucional formal; depende de voluntad política
Posible Protocolo Adicional Budapest sobre IA	Miembros Budapest + adherentes	Propuesta doctrinal; no iniciada	Actualización tipológica específica; mecanismo ágil	Lentitud negociación; excluye China y Rusia

Elaboración propia a partir de Consejo de Europa (2022); Sieber (2020); Palazzi (2022); OCDE (2023); Comisión Europea (2024); Europol (2023).

El rol de los actores del ecosistema tecnológico en la reforma

La responsabilidad de las empresas tecnológicas en la configuración normativa

La reforma legislativa del derecho penal ante la IA no puede ser concebida como un proceso unilateral de decisión pública sin implicación de los actores privados del ecosistema tecnológico. Las empresas desarrolladoras y desplegadoras de sistemas de IA son simultáneamente los actores con mayor capacidad técnica para comprender los riesgos que sus sistemas generan, los actores con mayor influencia sobre el diseño de las soluciones técnicas que pueden mitigar esos riesgos y los actores con mayores incentivos económicos para resistir reformas que incrementen sus obligaciones de diligencia y su exposición a la responsabilidad penal. Esta posición ambivalente hace que la política de participación de las empresas tecnológicas en los procesos legislativos sea uno de los desafíos de gobernanza más delicados de la reforma (Pasquale, 2020; Hildebrandt, 2020; Zuboff, 2021).

La doctrina comparada ha identificado tres modelos de participación del sector privado en el proceso normativo. El modelo de la autorregulación pura, en el que las empresas adoptan voluntariamente códigos de conducta, estándares técnicos y mecanismos de auditoría interna sin imposición normativa externa, ha demostrado ser insuficiente para garantizar la protección de los derechos fundamentales en el ámbito de los crímenes algorítmicos; los fracasos documentados de los sistemas de autorregulación en el ámbito de la privacidad digital y de los sesgos algorítmicos ilustran los límites de este modelo (O'Neil, 2016; Eubanks, 2018).

El modelo de la corregulación, en el que el legislador establece los objetivos normativos y los principios de diseño, y las empresas colaboran en la elaboración de los estándares técnicos de implementación, es el que ha mostrado mayor efectividad en otros ámbitos de la regulación tecnológica y es el que el Reglamento de IA europeo adopta de manera predominante. El modelo de la regulación impuesta, en el que el legislador determina unilateralmente tanto los objetivos como los medios de implementación, puede ser necesario en los supuestos de riesgo más elevado, pero requiere capacidades técnicas de supervisión que muchas administraciones públicas no tienen actualmente (Comisión Europea, 2024; Floridi et al., 2022).

La participación de la sociedad civil y de los grupos afectados

Un elemento determinante de la legitimidad democrática de las reformas legislativas del derecho penal ante la IA es la participación de los grupos de la sociedad civil que se ven más directamente afectados por los riesgos de los crímenes algorítmicos y de la justicia penal automatizada. Entre estos grupos, la literatura destaca con especial énfasis las organizaciones que representan a las comunidades históricamente más afectadas por los sesgos del sistema de justicia penal, comunidades afrodescendientes, indígenas, inmigrantes, personas en situación de pobreza, y que son las más vulnerables a los impactos diferenciados de los algoritmos de evaluación de riesgo y de los sistemas de reconocimiento facial biométrico (Angwin et al., 2016; Buolamwini & Gebru, 2020; Richardson et al., 2019).

La participación efectiva de estos grupos en los procesos legislativos requiere medidas específicas. La accesibilidad de la información técnica, traducida a lenguaje comprensible para los no especialistas, es una condición previa indispensable. La financiación pública de organizaciones de la sociedad civil que representan a los grupos más afectados permite que estas participen con recursos suficientes en los procesos de consulta. La adopción de mecanismos de consulta previa, libre e informada respecto a los pueblos indígenas en los países latinoamericanos donde el Convenio 169 de la OIT es de aplicación, es un requisito de derecho internacional específico que los procesos legislativos deben respetar cuando los sistemas de IA afecten a las comunidades indígenas (CIDH, 2021; Ortega & Salinas, 2023).

DISCUSIÓN Y CONCLUSIONES

Discusión de los hallazgos

Los resultados expuestos en las secciones precedentes permiten construir una visión de conjunto sobre el estado de la reforma del derecho penal ante la IA y las líneas de acción prioritarias que los legisladores, los operadores jurídicos y los actores del ecosistema tecnológico deberían seguir. La discusión de estos hallazgos debe situarse en el contexto más amplio del debate sobre la gobernanza de las tecnologías emergentes y su relación con el Estado de Derecho.

El hallazgo más relevante y preocupante de la revisión es la constatación de que el desfase normativo global entre el derecho penal vigente y los desafíos de los crímenes algorítmicos no solo persiste, sino que se está ampliando. La velocidad de adopción de los sistemas de IA

en los ámbitos de la seguridad, las finanzas, las comunicaciones y la infraestructura crítica supera con creces la velocidad de los procesos legislativos ordinarios, y la brecha resultante genera zonas de impunidad que los actores más sofisticados del ecosistema criminal explotan de manera sistemática. Paradójicamente, los sistemas más vulnerables al forum shopping punitivo algorítmico son los de los países con menor densidad normativa, precisamente los países de menor desarrollo normativo, lo que convierte la fragmentación normativa global en un mecanismo de exportación de riesgos desde las jurisdicciones más reguladas hacia las menos reguladas (Sieber, 2020; Palazzi, 2022; Europol, 2023).

El segundo hallazgo relevante es la tensión entre la urgencia de la reforma y la necesidad de reflexión deliberativa. Varios actores del debate argumentan que la urgencia de los riesgos que los crímenes algorítmicos representan justifica la adopción de reformas legislativas aceleradas sin la deliberación y la consulta que los principios del Estado de Derecho normalmente exigen. Esta argumentación debe ser resistida con energía, por dos razones. La primera es que las reformas adoptadas sin deliberación suficiente tienden a ser más fácilmente capturadas por los intereses de los actores más poderosos, en este caso, las empresas tecnológicas y las agencias de seguridad, y a reproducir o amplificar las desigualdades estructurales del sistema de justicia penal en lugar de corregirlas. La segunda razón es que las reformas penales apresuradas, si resultan incompatibles con los estándares constitucionales y de derechos humanos, pueden ser anuladas por los tribunales, generando mayor inseguridad jurídica que la que pretendían corregir. El equilibrio entre urgencia y deliberación es uno de los problemas de gobernanza más delicados de la reforma del derecho penal ante la IA (Hildebrandt, 2020; Floridi et al., 2022; CIDH, 2021).

El tercer hallazgo relevante es la centralidad del principio de proporcionalidad como criterio orientador de la reforma. El derecho penal es el instrumento de control social de mayor intensidad coercitiva disponible para el Estado, y su uso debe reservarse para las conductas de mayor gravedad que no pueden ser adecuadamente abordadas por otros mecanismos jurídicos menos intrusivos. En el ámbito de los crímenes algorítmicos, el principio de proporcionalidad exige que el legislador evalúe cuidadosamente si la creación de nuevos tipos penales o el incremento de las penas existentes es realmente necesario para la protección de los bienes jurídicos en riesgo, o si mecanismos de derecho administrativo sancionador, de

responsabilidad civil o de regulación técnica pueden alcanzar los mismos objetivos con menor coste para los derechos fundamentales (Mir Puig, 2023; Roxin, 2020; Ferrajoli, 2011).

En relación con los trabajos precedentes, los hallazgos del presente capítulo confirman y desarrollan las propuestas formuladas por Sieber (2020) para la reforma sistémica del derecho penal internacional de la tecnología, enriquecen el análisis comparado de Clough (2021) con los desarrollos normativos más recientes en la UE y América Latina, y actualizan las propuestas de Dremliuga & Kucherov (2022) sobre la tipificación de los crímenes algorítmicos con las lecciones de los primeros procesos legislativos específicos (DEFIANCE Act, Reglamento UE IA, Ley chilena 21.459). El presente análisis también confronta críticamente la posición de Hallevy (2022) sobre la personalidad jurídica electrónica, coincidiendo con la doctrina mayoritaria en que su prematura adopción en el ámbito penal generaría más problemas de los que resolvería, dado el estado actual del desarrollo tecnológico y los principios garantistas irrenunciables del derecho penal liberal.

Conclusiones

De la investigación desarrollada en el presente capítulo y en el conjunto de la obra de la que forma parte es posible extraer las siguientes conclusiones, formuladas con la cautela epistemológica que impone la naturaleza prospectiva de parte del análisis y el estado de acelerada evolución del campo.

La adaptación del derecho penal frente a la IA no es una opción de política legislativa disponible, sino una obligación jurídica del Estado de Derecho. Los derechos fundamentales de las víctimas de crímenes algorítmicos, a la tutela judicial efectiva, a la reparación, a la no discriminación, generan una obligación positiva de los Estados de articular marcos normativos efectivos que garanticen su protección. La impunidad estructural de los crímenes algorítmicos derivada del desfase normativo no es un resultado inevitable de la innovación tecnológica, sino una omisión legislativa que compromete la legitimidad del sistema de justicia penal.

La reforma debe ser sistémica, no cosmética. La simple agravación de los tipos existentes o la aplicación extensiva de las categorías doctrinales clásicas no es suficiente para afrontar los desafíos cualitativamente nuevos que los crímenes algorítmicos plantean. Se requiere una reforma que articule simultáneamente la actualización de los tipos penales, la reformulación de los criterios de imputación, la modernización del estatuto procesal de la evidencia

algorítmica, el fortalecimiento de las capacidades institucionales de investigación y persecución, y la construcción de marcos internacionales de cooperación específicamente adaptados a la transnacionalidad de estos crímenes.

La reforma debe preservar los principios garantistas del derecho penal liberal. El principio de culpabilidad, el principio de legalidad, el principio de proporcionalidad y el principio de presunción de inocencia no son obstáculos para la reforma, sino el marco normativo dentro del cual la reforma debe operar. Cualquier adaptación del derecho penal ante la IA que sacrifique estos principios no fortalece el Estado de Derecho, sino que lo erosiona. La reforma debe buscar nuevas soluciones dogmáticas, la imputación diferenciada por niveles de autonomía, la figura del garante digital, la responsabilidad corporativa por defecto de organización algorítmica, que cubran el vacío de responsabilidad sin abandonar los principios que dotan de legitimidad al sistema punitivo.

La cooperación internacional es indispensable y urgente. La transnacionalidad estructural de los crímenes algorítmicos hace que ningún ordenamiento jurídico nacional pueda afrontar estos desafíos de manera aislada. La adhesión al Convenio de Budapest, la participación activa en los procesos de elaboración de la Convención de la ONU y el impulso de iniciativas regionales de armonización normativa, especialmente en América Latina, son prioridades de política exterior y de seguridad que los Estados deben abordar con la urgencia que la situación demanda.

El proceso de reforma debe ser inclusivo y deliberativo. La legitimidad democrática de las reformas legislativas en el ámbito del derecho penal ante la IA depende de que los grupos más vulnerables a los impactos diferenciados de los algoritmos sean genuinamente escuchados en el proceso. La reforma sin participación efectiva de estos grupos reproducirá las desigualdades estructurales del sistema de justicia con el ropaje de la objetividad técnica.

En cuanto a las limitaciones del presente estudio, cabe señalar que la naturaleza prospectiva del análisis implica un nivel de incertidumbre inherente que ninguna metodología puede eliminar completamente. Las líneas de investigación futura que se derivan de los hallazgos incluyen: el análisis empírico del impacto de las reformas legislativas ya adoptadas sobre la tasa de persecución penal de los crímenes algorítmicos; el estudio comparado de los mecanismos de implementación de la responsabilidad corporativa penal en el ámbito de la IA

en distintos sistemas jurídicos; la evaluación del impacto diferencial de los algoritmos de predicción de riesgo sobre grupos étnicos y socioeconómicos en contextos latinoamericanos mediante análisis de datos judiciales; y el desarrollo de propuestas dogmáticas específicas para la imputación de responsabilidad en los supuestos de daño producido por sistemas de IA de alta autonomía en los ordenamientos jurídicos de América Latina.

CONCLUSIONES FINALES

I. Consideración Preliminar

La presente obra ha recorrido de manera sistemática y progresiva uno de los desafíos jurídicos más complejos y urgentes de nuestra época: la intersección entre la inteligencia artificial y el derecho penal. A lo largo de sus seis capítulos, el libro ha trazado un itinerario analítico que va desde los fundamentos conceptuales hasta las propuestas de reforma legislativa, pasando por los problemas de responsabilidad, las nuevas tipologías delictivas, la evidencia algorítmica y la dimensión ética y de derechos humanos. Las conclusiones finales que siguen no pretenden reiterar los resultados específicos de cada capítulo, sino sintetizar los hallazgos transversales de la obra en su conjunto, articulando las tesis centrales que emergen de su lectura integrada y las líneas de acción que de ellas se derivan.

II. El desfase estructural como punto de partida irrenunciable

La conclusión más robusta y transversal de la obra es la constatación de un desfase estructural entre el derecho penal vigente y los sistemas de inteligencia artificial. Este desfase no es cuantitativo (el derecho simplemente va rezagado respecto a la tecnología), sino cualitativo: las categorías fundamentales del derecho penal (acción, culpabilidad, causalidad, autoría, prueba) fueron construidas sobre el supuesto de un sujeto activo humano, consciente, voluntario e identificable, y no disponen de instrumentos adecuados para responder a los agentes algorítmicos autónomos, opacos y multi-actor que la inteligencia artificial ha introducido en el mundo social.

Este desfase se manifiesta en cuatro dimensiones que la obra ha documentado con detalle: la dimensión tipológica (los crímenes algorítmicos de tercera generación carecen de cobertura penal satisfactoria), la dimensión de la imputación (el vacío de responsabilidad generado por los sistemas de alta autonomía no ha sido resuelto por ningún ordenamiento), la dimensión probatoria (la evidencia algorítmica carece de un estatuto procesal adecuado en la mayoría

de los sistemas jurídicos), y la dimensión de la cooperación internacional (la territorialidad del derecho penal resulta estructuralmente inadecuada frente a la transnacionalidad constitutiva de los crímenes algorítmicos).

El reconocimiento de este desfase como punto de partida no es un acto de pesimismo intelectual, sino la condición necesaria de una reforma honesta y eficaz. Solo quien comprende la profundidad y la naturaleza del problema puede aspirar a respuestas normativas a la altura del desafío.

III. Los principios garantistas como límites infranqueables y como guías de la reforma

Una segunda conclusión central de la obra es que los principios cardinales del derecho penal liberal (legalidad, culpabilidad, presunción de inocencia, proporcionalidad y tutela judicial efectiva) no son obstáculos para la adaptación del sistema punitivo ante la inteligencia artificial, sino el marco normativo dentro del cual esa adaptación debe producirse. Cualquier reforma que sacrifique estos principios en aras de la eficiencia técnica o de la urgencia de la respuesta no fortalece el Estado de Derecho; lo erosiona.

El principio de culpabilidad, en particular, opera como límite constitucional infranqueable de los modelos de imputación penal ante sistemas autónomos de IA. Los modelos que postulan la responsabilidad objetiva plena o la personalidad jurídico-penal del sistema de IA son incompatibles con este principio y no deben adoptarse sin una reforma constitucional previa que modifique los fundamentos filosóficos del ius puniendi, modificación que sería filosóficamente cuestionable en el estado actual del conocimiento sobre la naturaleza de la IA.

El principio de legalidad, por su parte, impone exigencias específicas al proceso de tipificación de los crímenes algorítmicos: los tipos penales deben ser lo suficientemente precisos para que sus destinatarios puedan prever las consecuencias jurídicas de sus conductas, sin que esa precisión implique una rigidez que los haga obsoletos ante la aceleración tecnológica. Las técnicas legislativas de cláusulas abiertas con remisión a estándares técnicos actualizables, de tipos de peligro abstracto funcionales y de mecanismos de revisión periódica, son las respuestas más adecuadas a esta tensión.

El principio de presunción de inocencia prohíbe que las decisiones penales relevantes (especialmente la privación de libertad preventiva) se fundamenten en evaluaciones

estadísticas grupales que traten al individuo como representante de su grupo demográfico antes que como persona con sus propias circunstancias y responsabilidades. Los sistemas actuariales de evaluación de riesgo que no respetan esta exigencia son incompatibles con los estándares del Pacto Internacional de Derechos Civiles y Políticos y de la Convención Americana sobre Derechos Humanos.

IV. El vacío de responsabilidad: un problema real pero soluble

La obra ha documentado con rigor la existencia de un vacío de responsabilidad penal en los supuestos de daño producido por sistemas de IA con elevada autonomía. Este vacío resulta de la combinación de tres factores que se potencian mutuamente: la distancia causal entre la conducta humana y el resultado (mediada por procesos algorítmicos que el propio creador del sistema no puede predecir con exactitud), la opacidad del proceso decisorio interno (el problema de la caja negra), y la distribución de tareas entre múltiples actores en la cadena de desarrollo y despliegue que diluye la imputabilidad individual.

Sin embargo, la obra concluye también que este vacío no es total ni absoluto: se produce principalmente en los supuestos de comportamiento emergente genuinamente imprevisible de sistemas de alta autonomía, que en el estado actual del desarrollo tecnológico constituyen una minoría de los casos. Para la mayoría de supuestos de daño producido por sistemas de IA de mediana y baja autonomía, los modelos dogmáticos existentes (imputación distribuida entre los actores de la cadena, comisión por omisión del garante digital, responsabilidad corporativa por defecto de organización) ofrecen instrumentos conceptuales suficientes, siempre que se realice el esfuerzo interpretativo de actualizarlos al nuevo contexto tecnológico.

La superación del vacío en los supuestos de alta autonomía requiere reformas legislativas específicas: la codificación de los deberes de diligencia exigibles a cada actor de la cadena de IA con referencia a estándares técnicos certificados; la extensión de la responsabilidad penal corporativa al ámbito del desarrollo y despliegue de sistemas de IA de alto riesgo, sobre el fundamento del defecto de organización; y, subsidiariamente, la creación de fondos sectoriales de compensación que garanticen la reparación a las víctimas en los supuestos de impunidad individual inevitables desde la perspectiva del principio de culpabilidad.

V. Los crímenes algorítmicos: una nueva categoría que demanda respuesta normativa urgente

La obra ha cartografiado la tipología emergente de los crímenes algorítmicos de tercera generación: el fraude por IA generativa (incluyendo el Business Email Compromise potenciado por clonación de voz y la manipulación automatizada de mercados financieros), los deepfakes delictivos (en sus dimensiones de desinformación electoral, fraude de identidad sintética, extorsión por contenido sexual no consentido y manipulación de evidencia procesal), los ciberataques autónomos contra infraestructuras críticas, la manipulación de sistemas de decisión automatizada mediante ataques adversariales y envenenamiento de datos, los crímenes de datos mediados por vigilancia algorítmica, y la desinformación algorítmica orquestada como ataque a los procesos democráticos.

Estas modalidades comparten características que las distinguen cualitativamente de los delitos informáticos clásicos: la escala potencial de los ataques, que puede afectar simultáneamente a millones de víctimas; la sofisticación técnica, que supera la capacidad de detección de los mecanismos de seguridad convencionales; la velocidad de ejecución, que supera la capacidad de respuesta humana; y la dificultad de atribución, que hace prácticamente imposible identificar al responsable mediante las técnicas forenses ordinarias.

La respuesta normativa a estos crímenes requiere una estrategia mixta que combine la actualización interpretativa de los tipos penales existentes con la tipificación específica de las modalidades más lesivas y más claramente diferenciadas; la creación de mecanismos de cooperación internacional específicamente adaptados a la transnacionalidad de estas conductas; y el fortalecimiento de las capacidades institucionales de investigación y persecución mediante la creación de unidades especializadas con competencias técnicas interdisciplinares.

VI. La evidencia algorítmica: el eslabón más débil del sistema de justicia

Uno de los hallazgos más perturbadores de la obra es la constatación de que los sistemas de justicia penal están utilizando de manera rutinaria evidencia generada por sistemas de IA sin disponer de marcos normativos adecuados que garanticen su fiabilidad, su controlabilidad y su compatibilidad con las garantías del proceso equitativo. Los sistemas de evaluación de riesgo de reincidencia, los sistemas de reconocimiento facial, las herramientas de análisis de

datos masivos y los sistemas de análisis forense automatizado producen conclusiones que determinan decisiones sobre la libertad de las personas sin que los acusados tengan acceso real a la información necesaria para cuestionarlas.

La evidencia algorítmica requiere un estatuto probatorio específico articulado en torno a cuatro principios irrenunciables: el principio de validación técnica previa a la admisión (el sistema debe haber sido validado de manera independiente antes de que sus conclusiones puedan utilizarse como prueba de cargo); el principio de auditabilidad (la defensa tiene derecho a que peritos de su parte puedan examinar el sistema y sus condiciones de funcionamiento); el principio de explicabilidad local (el sistema debe poder proporcionar una explicación comprensible de los factores que determinaron su conclusión en el caso concreto); y el principio de igualdad de armas técnica (los sistemas de justicia deben garantizar a la defensa acceso a peritos especializados y recursos técnicos suficientes para ejercer efectivamente el derecho de contradicción).

El sesgo algorítmico documentado en los sistemas más utilizados (tasas de error hasta dos veces mayores para imputados de raza negra que para imputados de raza blanca en los sistemas de evaluación de riesgo; tasas de error hasta treinta veces mayores en el reconocimiento de mujeres de piel oscura que en el de hombres de piel clara en los sistemas biométricos) no es un defecto técnico accidental, sino la expresión matemática de inequidades históricas profundamente arraigadas en los datos de entrenamiento. Su corrección requiere tanto medidas técnicas de desbiasado como mecanismos jurídicos que permitan la impugnación efectiva de las decisiones fundadas en sistemas con sesgo discriminatorio documentado.

VII. La dimensión ética: cuando la automatización reproduce y amplifica la desigualdad

La obra ha demostrado que la automatización de la justicia penal no es éticamente neutra. Los sistemas de IA no eliminan los sesgos del juzgador humano, sino que codifican y automatizan los sesgos presentes en los datos históricos sobre los que son entrenados, confiriéndoles una apariencia de objetividad técnica que los hace más difíciles de identificar e impugnar. La promesa de objetividad del algoritmo es, paradójicamente, uno de sus riesgos más graves.

Los derechos a la no discriminación, la dignidad humana, el debido proceso, la presunción de inocencia y la privacidad son sistemáticamente amenazados por el uso no controlado de sistemas algorítmicos en la justicia penal. La evidencia empírica acumulada durante la última década demuestra que estas amenazas no son hipotéticas sino actuales y documentadas, con consecuencias reales sobre vidas humanas concretas y sobre grupos que ya sufrían las desigualdades del sistema de justicia penal antes de la introducción de los algoritmos.

La compatibilidad de la justicia penal automatizada con el Estado de Derecho exige marcos éticos vinculantes (no meramente declarativos), mecanismos de rendición de cuentas algorítmica (que asignen responsabilidad a todos los actores del ecosistema de la IA en función de su posición en la cadena de desarrollo y despliegue), y garantías de supervisión humana significativa (no meramente formal). El principio de no delegación de decisiones irreversibles a sistemas algorítmicos debe ser consagrado legislativamente: ninguna decisión que implique privación de libertad puede adoptarse exclusivamente sobre la base de conclusiones algorítmicas.

VIII. La urgencia de la reforma en perspectiva comparada

La revisión comparada de los marcos normativos vigentes revela que ningún ordenamiento jurídico ha articulado aún una respuesta sistémica y coherente a los desafíos que la IA plantea al derecho penal. El sistema europeo, con el Reglamento de IA (2024/1689) como instrumento más avanzado a nivel global, ha dado el paso más significativo al adoptar un marco jurídicamente vinculante de carácter preventivo, pero carece todavía de tipos penales específicos para los crímenes algorítmicos y de un estatuto procesal armonizado para la evidencia algorítmica. El sistema estadounidense es marcadamente fragmentado, con respuestas sectoriales que no se articulan en una arquitectura normativa coherente. Los países latinoamericanos presentan, en su mayoría, marcos de primera o segunda generación diseñados para el cibercrimen clásico que resultan manifiestamente insuficientes para los crímenes algorítmicos de tercera generación.

La situación latinoamericana merece una atención específica. La región se enfrenta a una doble vulnerabilidad: sus sistemas de justicia penal exhiben desigualdades estructurales más pronunciadas que las de las democracias avanzadas del Norte Global, lo que hace que los sesgos de los sistemas algorítmicos sean potencialmente más dañinos; y su capacidad

institucional de supervisión y auditoría de los sistemas de IA es menor, lo que hace más difícil el control efectivo de los sistemas desplegados. Paradójicamente, esta doble vulnerabilidad convierte la situación latinoamericana en una oportunidad singular: los países de la región pueden diseñar desde el principio marcos normativos que incorporen los estándares de garantía más exigentes, aprendiendo de los errores y las insuficiencias de los sistemas más avanzados, sin necesidad de pasar por el ciclo de reforma reactiva que ha caracterizado la respuesta de esos sistemas.

La adhesión progresiva al Convenio de Budapest sobre Ciberdelincuencia (ya completada por Chile y Colombia, en proceso en Brasil y Argentina, iniciada por Ecuador) es un primer paso necesario pero insuficiente. Debe ser complementada con la adopción de legislaciones específicas que aborden los crímenes algorítmicos de manera sustantiva, con el fortalecimiento de las capacidades técnicas de las instituciones del sistema de justicia, con la creación de organismos de auditoría independiente, y con el impulso de iniciativas regionales de coordinación normativa y cooperación institucional.

IX. Los cuatro ejes indispensables de la reforma

La obra ha identificado cuatro ejes de reforma que son indispensables y complementarios entre sí para articular una respuesta sistémica adecuada a los desafíos que la IA plantea al derecho penal.

El primer eje es la actualización de los tipos penales. Esta actualización debe seguir una estrategia mixta: la introducción de circunstancias agravantes por el uso de IA en los tipos existentes (para conductas que la IA amplifica cuantitativamente pero no transforma cualitativamente), combinada con la tipificación autónoma de las conductas que presentan una especificidad técnica y jurídica propia (para crímenes como el envenenamiento de datos de entrenamiento, la creación de agentes de IA maliciosos con finalidad delictiva, la desinformación algorítmica orquestada y la manipulación de sistemas de decisión de interés público). La técnica legislativa de las cláusulas abiertas con remisión a estándares técnicos actualizables permite conciliar la precisión del principio de legalidad con la flexibilidad que la dinámica tecnológica impone.

El segundo eje es la reformulación de los criterios de imputación. El modelo de imputación diferenciada por niveles de autonomía del sistema, combinado con la codificación de los

deberes de diligencia exigibles a cada actor de la cadena (desarrollador, fabricante, desplegador, usuario) y con la extensión de la responsabilidad corporativa penal por defecto de organización al ámbito de la IA de alto riesgo, es el instrumento dogmático más compatible con los principios garantistas del derecho penal liberal y, al mismo tiempo, el que mayor aptitud demuestra para cerrar el vacío de responsabilidad documentado.

El tercer eje es la modernización del estatuto procesal de la evidencia algorítmica. La construcción de un estatuto probatorio específico para la evidencia algorítmica, articulado en torno a los principios de validación técnica, auditabilidad, explicabilidad local e igualdad de armas técnica, es una tarea normativa de primera urgencia que ningún ordenamiento ha completado de manera satisfactoria. Este estatuto debe complementarse con la modernización de los protocolos de cadena de custodia para la evidencia digital y con la creación de organismos de auditoría forense algorítmica independientes con los recursos técnicos y jurídicos necesarios.

El cuarto eje es la construcción de marcos supranacionales de cooperación. La transnacionalidad estructural de los crímenes algorítmicos requiere mecanismos de cooperación internacional específicamente adaptados a sus características: la velocidad de los ataques, la dispersión geográfica de la infraestructura, las dificultades de atribución y la opacidad de los sistemas utilizados. La Convención de la ONU sobre Ciberdelincuencia de 2024 es un primer paso en esta dirección, que debe ser complementado con la elaboración de un Protocolo Internacional sobre Crímenes Algorítmicos en el marco del Convenio de Budapest, con el impulso de iniciativas regionales de armonización normativa (especialmente en América Latina), y con programas robustos de fortalecimiento de capacidades técnicas en los países con menor desarrollo institucional.

X. El proceso de reforma: condiciones de legitimidad

La obra concluye que la reforma del derecho penal ante la IA no puede limitarse a sus contenidos sustantivos, sino que debe incorporar condiciones de proceso que garanticen su legitimidad democrática. Tres condiciones son especialmente relevantes.

La primera condición es la deliberación inclusiva. La reforma legislativa del derecho penal ante la IA afecta directamente a los grupos que ya experimentan las mayores desigualdades del sistema de justicia penal, y su legitimidad depende de que esos grupos sean genuinamente

escuchados en el proceso. La participación efectiva requiere medidas específicas: la accesibilidad de la información técnica en lenguaje comprensible para los no especialistas, la financiación pública de organizaciones de la sociedad civil representativas de los grupos más afectados, y la adopción de mecanismos de consulta previa, libre e informada donde el derecho internacional lo exige.

La segunda condición es la interdisciplinariedad. Los problemas que la IA plantea al derecho penal no pueden ser resueltos desde ninguna disciplina aislada. La dogmática jurídico-penal, la ciencia de la computación, la filosofía del derecho, la ética aplicada y la criminología deben converger en un esfuerzo analítico conjunto que supere las fronteras disciplinares y produzca respuestas normativas técnicamente informadas y jurídicamente sólidas. La formación interdisciplinaria de los operadores jurídicos (jueces, fiscales, abogados defensores, legisladores) es una condición estructural de la reforma que los sistemas de educación jurídica no pueden seguir ignorando.

La tercera condición es la proporcionalidad. El derecho penal es el instrumento de control social de mayor intensidad coercitiva disponible para el Estado, y su uso debe reservarse para las conductas de mayor gravedad. Frente a la presión de ampliar el ámbito penal en respuesta a los nuevos riesgos tecnológicos, la reforma debe aplicar un principio de mínima intervención penal que evalúe cuidadosamente si la creación de nuevos tipos o el incremento de las penas existentes es realmente necesario, o si mecanismos de derecho administrativo sancionador, de responsabilidad civil o de regulación técnica pueden alcanzar los mismos objetivos con menor coste para los derechos fundamentales.

XI. Reflexión final: la ia como espejo del derecho penal

Existe una dimensión de la relación entre la inteligencia artificial y el derecho penal que trasciende los problemas técnicos y normativos específicos que la obra ha analizado: la IA actúa como un espejo que devuelve al derecho penal una imagen amplificada de sus propias contradicciones y desigualdades históricas.

Los sesgos de los sistemas algorítmicos no son ajenos al derecho penal: son los sesgos del propio sistema de justicia, codificados y automatizados. Los vacíos de responsabilidad que la IA genera no son enteramente nuevos: son la versión contemporánea de los problemas clásicos de imputación en organizaciones complejas, que la IA hace más visibles al

exacerbarlos. Y los déficits de transparencia de la evidencia algorítmica no son una anomalía técnica: reflejan la opacidad de un sistema de justicia que históricamente ha concedido a determinados tipos de prueba pericial un crédito epistémico superior al que su verificabilidad justificaba.

Esta dimensión reflexiva de la IA como espejo del derecho penal no es un motivo de parálisis, sino de esperanza. Si la irrupción de la inteligencia artificial nos obliga a reexaminar críticamente los presupuestos y las prácticas del sistema punitivo, a confrontar sus desigualdades históricas, a exigir mayor transparencia y mayor rendición de cuentas, y a construir garantías más robustas para los más vulnerables, entonces la crisis que plantea puede convertirse en la palanca de una reforma que el derecho penal necesitaba mucho antes de que los algoritmos llegaran.

La tarea que esta obra ha intentado iniciar, la de contribuir a una dogmática penal tecnológicamente informada que sea capaz de preservar las garantías fundamentales del individuo frente al poder punitivo del Estado en el entorno de la sociedad digital, es una tarea que corresponde a toda la comunidad jurídica: legisladores, jueces, fiscales, abogados, académicos y, sobre todo, a quienes sufren en sus propias vidas las consecuencias de un sistema de justicia todavía incompleto. Este libro es una contribución a ese esfuerzo colectivo, con la conciencia de que el ritmo de los cambios tecnológicos exigirá continuar y profundizar esa contribución de manera sistemática, rigurosa y comprometida con la dignidad de todas las personas.

REFERENCIAS BIBLIOGRAFICAS

Access Now. (2023). Analysis of the UN Cybercrime Convention text. Access Now Policy Brief. https://accessnow.org/un-cybercrime-convention-analysis

Alto Comisionado de las Naciones Unidas para los Derechos Humanos [ACNUDH]. (2021). The right to privacy in the digital age: Artificial intelligence and human rights. United Nations. https://doi.org/10.18356/9789210058612

Alto Comisionado de las Naciones Unidas para los Derechos Humanos [ACNUDH]. (2023). Impact of artificial intelligence on the enjoyment of human rights. A/HRC/53/40. United Nations. https://www.ohchr.org/en/documents/thematic-reports/ahrc5340-impact-artificial-intelligence-enjoyment-human-rights

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. ProPublica. Recuperado de https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012

Berk, R. A. (2021). Artificial Intelligence, Predictive Policing, and Risk Assessment for Law Enforcement. Annual Review of Criminology, 4(1), 209-237. https://doi.org/10.1146/annurev-criminol-051520-012342

Biggio, B., & Roli, F. (2022). Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition, 84, 317–331. https://doi.org/10.1016/j.patcog.2018.07.023

Binns, R. (2022). Algorithmic accountability and public reason. Philosophy & Technology, 35(1), 1–26. https://doi.org/10.1007/s13347-018-0334-6

Binns, R. (2022). Human judgment in algorithmic loops: Individual justice and automated decision-making. Regulation & Governance, 16(1), 197-211. https://doi.org/10.1111/rego.12358

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., Farquhar, S., & Amodei, D. (2021). The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. arXiv. https://doi.org/10.48550/arXiv.1802.07228

Buolamwini, J., & Gebru, T. (2020). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of Machine Learning Research, 81, 77-91. https://proceedings.mlr.press/v81/buolamwini18a.html

Calo, R. (2020). Artificial intelligence policy: A primer and roadmap. UC Davis Law Review, 51(2), 399-435.

Canca, C. (2020). Operationalizing AI ethics principles. Communications of the ACM, 63(12), 18-21. https://doi.org/10.1145/3430368

Cannataci, J. A. (2021). Report of the Special Rapporteur on the right to privacy: Artificial intelligence and privacy. A/HRC/46/37. United Nations Human Rights Council. https://www.ohchr.org/en/issues/privacy/pages/reportsfhrc.aspx

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., & Raffel, C. (2021). Extracting training data from large language models. Proceedings of the 30th USENIX Security Symposium, pp. 2633–2650.

Casey, E. (2019). Handbook of digital forensics and investigation (2nd ed.). Academic Press, pp. 1–594.

Casey, E. (2019). The chilling effect of uncertainty on digital forensics. Digital Investigation, 28, A1-A3. https://doi.org/10.1016/j.diin.2019.01.001

Casey, E. (2019). Digital evidence and computer crime: Forensic science, computers and the internet (4th ed.). Academic Press.

Char, D. S., Shah, N. H., & Magnus, D. (2020). Implementing machine learning in health care—addressing ethical challenges. New England Journal of Medicine, 378(10), 981–983. https://doi.org/10.1056/NEJMp1714229

Cheng, A., Farber, B., & Nicoson, R. (2024). Evaluating algorithmic evidence in criminal courts: Toward a unified admissibility standard. Journal of Criminal Law & Criminology, 114(2), 289–345. https://doi.org/10.2307/jcrimlawcriminology.114.2.289

Cheng, E., Barry, M., & Nunn, K. (2024). Algorithmic evidence: Admissibility, reliability and the future of criminal trials. Stanford Technology Law Review, 27(1), 1–68. https://doi.org/10.2139/ssrn.4521847

Chesney, R., & Citron, D. K. (2022). Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. California Law Review, 107(6), 1753-1820. https://doi.org/10.15779/Z38RV0D15J

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), 153-163. https://doi.org/10.1089/big.2016.0047

Clough, J. (2021). Principles of cybercrime (3rd ed.). Cambridge University Press, pp. 1–530. https://doi.org/10.1017/9781108596312

Comisión Europea. (2023). Digital finance package: Regulation on markets in crypto-assets (MiCA). Official Journal of the European Union, L 150/40.

Comisión Europea. (2024). Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial. Diario Oficial de la Unión Europea. https://doi.org/10.32657/10356/169143

Comisión Interamericana de Derechos Humanos [CIDH]. (2021). Informe temático: Derechos humanos de las personas migrantes y otras en el contexto de la movilidad humana y tecnología. CIDH. https://www.oas.org/es/cidh/informes/tematicos.asp

Comisión Interamericana de Derechos Humanos. (2021). Informe temático sobre tecnología y derechos humanos. OEA. http://www.oas.org/es/cidh/informes/tematicos.asp

Comisión Europea. (2024). Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo relativo a la inteligencia artificial (Reglamento de Inteligencia Artificial). Diario Oficial de la Unión Europea, L 2024/1689. https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=OJ:L_202401689

Consejo de Derechos Humanos [CDH]. (2023). Resolución 53/16: Nuevas tecnologías y derechos humanos. A/HRC/RES/53/16. Naciones Unidas. https://www.ohchr.org/en/hr-bodies/hrc/sessions/53/resolutions

Consejo de Europa. (2020). Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems. Council of Europe. https://rm.coe.int/1680a0c08d

Consejo de Europa. (2001). Convention on Cybercrime (Budapest Convention), ETS No. 185. Council of Europe Treaty Series.

Consejo de Europa. (2022). Second Additional Protocol to the Convention on Cybercrime on enhanced co-operation and disclosure of electronic evidence, CETS No. 224.

Cummings, M. L. (2017). Automation bias in intelligent time critical decision support systems. AIAA 1st Intelligent Systems Technical Conference. https://doi.org/10.2514/6.2004-6313

Department of Justice, U.S. (2023). National cybersecurity strategy: Department of Justice implementation plan. U.S. Department of Justice.

Doshi-Velez, F., & Kim, B. (2021). Towards a rigorous science of interpretable machine learning. arXiv. https://doi.org/10.48550/arXiv.1702.08608

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., Scott, K., Schieber, S., Waldo, J., Weinberger, D., Weller, A., & Wood, A. (2021). Accountability of AI under the law: The role of explanation. Berkman Klein Center Research Publication. https://doi.org/10.2139/ssrn.3064761

Dremliuga, R., & Kucherov, I. (2022). Criminal Liability for Artificial Intelligence: Critical Analysis of the Russian Draft Law. Russian Law Journal, 10(2), 116-140. https://doi.org/10.17589/2309-8678-2022-10-2-116-140

Dremliuga, R., & Kucherov, A. (2022). Criteria for criminal liability of artificial intelligence systems. Kutafin Law Review, 9(1), 143–165. https://doi.org/10.17803/2313-5395.2022.1.19.143-165

Dremliuga, R., & Kucherov, I. (2022). Algorithmic crimes: Towards a new typology of criminal offenses. Computer Law & Security Review, 46, 105736. https://doi.org/10.1016/j.clsr.2022.105736

Dremliuga, R., & Kucherov, N. (2022). Artificial intelligence and criminal liability: An analysis of the legal problems of machine decision-making and autonomy. International Journal of Law and Information Technology, 30(2), 148–172. https://doi.org/10.1093/ijlit/eaac012

Dressel, J., & Farid, H. (2022). The accuracy, fairness, and limits of predicting recidivism. Science Advances, 4(1), eaao5580. https://doi.org/10.1126/sciadv.aao5580

Electronic Frontier Foundation (EFF). (2023). EFF's concerns about the UN Cybercrime Treaty. Electronic Frontier Foundation. https://www.eff.org/deeplinks/2023/04/eff-un-cybercrime-treaty

Eubanks, V. (2018). Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin's Press.

Europol. (2023). European Union serious and organised crime threat assessment (SOCTA 2023). Europol Publications Office, pp. 1–88. https://www.europol.europa.eu/publication-events/main-reports/socta-report

Europol. (2023). Europol report on cybercrime and AI. European Cybercrime Centre (EC3). https://www.europol.europa.eu/publications-events/publications/artificial-intelligence-how-europol-sees-use-of-ai-in-criminal-activities

Europol. (2023). Internet Organised Crime Threat Assessment (IOCTA) 2023. European Union Agency for Law Enforcement Cooperation. https://www.europol.europa.eu/publication-events/main-reports/internet-organised-crime-threat-assessment-iocta-2023

Farid, H. (2022). Creating, using, misusing, and detecting deep fakes. Journal of Online Trust and Safety, 1(4), 1–26. https://doi.org/10.54501/jots.v1i4.56

FBI. (2024). Internet crime report 2023. Federal Bureau of Investigation, Internet Crime Complaint Center (IC3), pp. 1–32.

Ferguson, A. G. (2021). Surveillance and the Tyrant Test. Georgetown Law Journal, 113(1), 1-76.

Ferguson, A. G. (2021). The rise of big data policing: Surveillance, race, and the future of law enforcement. New York University Press.

Ferrajoli, L. (2011). Derecho y razón: Teoría del garantismo penal (10ª ed.). Editorial Trotta.

Ferrara, E. (2023). The butterfly effect: How AI-enabled disinformation shapes political discourse. First Monday, 28(5). https://doi.org/10.5210/fm.v28i5.12957

Ferrara, E. (2023). The history of digital fake news. First Monday, 26(5). https://doi.org/10.5210/fm.v26i5.11969

FATF. (2024). Guidance on a risk-based approach to virtual assets and virtual asset service providers. Financial Action Task Force.

Floridi, L. (2021). The Ethics of Artificial Intelligence. Oxford University Press, pp. 1–350.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., & Vayena, E. (2022). An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Floridi, L., Cowls, J., King, T. C., & Taddeo, M. (2022). How to design AI for social good: Seven essential factors. Science and Engineering Ethics, 26(3), 1771-1796. https://doi.org/10.1007/s11948-020-00213-5

Gabriel, I. (2021). Towards a theory of justice for artificial intelligence. Daedalus, 151(2), 218–231. https://doi.org/10.1162/daed_a_01911

Garvie, C., Bedoya, A. M., & Frankle, J. (2022). The Perpetual Line-Up: Unregulated Police Face Recognition in America. Georgetown Law Center on Privacy and Technology. https://perpetuallineup.org

Gómez-Jara Díez, C. (2021). Responsabilidad penal de las personas jurídicas: Aspectos sustantivos y procesales (3.ª ed.). La Ley, pp. 1–420.

Gómez-Jara Díez, C. (2021). Corporate criminal liability in the digital economy: A comparative assessment. European Journal of Crime, Criminal Law and Criminal Justice, 29(2), 149–178. https://doi.org/10.1163/15718174-29020003

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. Minds and Machines, 30(1), 99-120. https://doi.org/10.1007/s11023-020-09517-8

Hallevy, G. (2022). When Robots Kill: Artificial Intelligence under Criminal Law. Northeastern University Press.

Hallevy, G. (2022). Liability for crimes involving artificial intelligence systems. Criminal Law and Philosophy, 16(1), 1–28. https://doi.org/10.1007/s11572-020-09528-7

Hallevy, G. (2022). A model of criminal responsibility for artificial intelligence: When criminals can be robots. Journal of International Commercial Law and Technology, 8(3), 192–212.

Hannah-Moffat, K. (2019). Actuarial sentencing: An unsettled proposition. Justice Quarterly, 36(2), 270-296. https://doi.org/10.1080/07418825.2018.1533343

Hassemer, W. (2018). Por qué no debe suprimirse el derecho penal subjetivo. Revista Penal México, 15, 41-57.

Henry, N., Flynn, A., & Powell, A. (2021). Image-based sexual abuse: Victims and perpetrators. Trends & Issues in Crime and Criminal Justice, 572, 1–19. https://doi.org/10.52922/ti78132

Hildebrandt, M. (2020). Law for Computer Scientists and Other Folk. Oxford University Press. https://doi.org/10.1093/oso/9780198860129.001.0001

Hildebrandt, M. (2020). Law for Computer Scientists and Other Folk. Oxford University Press, pp. 1–376.

Interpol. (2023). Interpol global crime trend summary 2023. INTERPOL General Secretariat, pp. 1–56. https://www.interpol.int/en/News-and-Events/News/2023/INTERPOL-report-charts-top-cyberthreats-in-Africa

IOSCO. (2022). Report on artificial intelligence and machine learning in securities markets. International Organization of Securities Commissions, IOSCO FR05/2022.

Jakobs, G. (2019). Derecho penal: Parte general. Fundamentos y teoría de la imputación (2.ª ed., trad. J. Cuello Contreras & J. Serrano González de Murillo). Marcial Pons, pp. 1–1113.

Jescheck, H. H., & Weigend, T. (2019). Tratado de derecho penal. Parte general (5.ª ed., trad. M. Olmedo Cardenete). Comares, pp. 1–986.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389-399. https://doi.org/10.1038/s42256-019-0088-2

Kant, I. (1785). Fundamentación de la metafísica de las costumbres (trad. M. García Morente). Tecnos (ed. 2007).

Karie, N. M., & Venter, H. S. (2020). Taxonomy of challenges for digital forensics. Journal of Forensic Sciences, 60(4), 885–893. https://doi.org/10.1111/1556-4029.12718

Kerr, O. S. (2020). Searches and seizures in a digital world. Harvard Law Review, 119(2), 531–585. https://doi.org/10.1017/9781316691854

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. The Quarterly Journal of Economics, 133(1), 237-293. https://doi.org/10.1093/qje/qjx032

LeCun, Y., Bengio, Y., & Hinton, G. (2021). Deep learning. Nature, 521(7553), 436-444. https://doi.org/10.1038/nature14539

Lin, H. S., & Zegart, A. (2019). Bytes, bombs, and spies: The strategic dimensions of offensive cyber operations. Brookings Institution Press, pp. 1–394.

Mata y Martín, R. M. (2021). Criminalidad informática: delitos cometidos a través de medios informáticos. Tirant lo Blanch, pp. 1–480.

Matthias, A. (2021). The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics and Information Technology, 6(3), 175-183. https://doi.org/10.1007/s10676-004-3422-1

Maras, M. H., & Alexandrou, A. (2023). Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos. International Journal of Evidence & Proof, 23(3), 255–262. https://doi.org/10.1177/1365712718807226

Metcalf, J., & Moss, E. (2019). Owning ethics: Corporate logics, Silicon Valley, and the institutionalization of AI ethics. Social Research: An International Quarterly, 86(2), 449-476. https://doi.org/10.1353/sor.2019.0029

Mir Puig, S. (2023). Derecho penal: Parte general (11.ª ed.). Reppertor, pp. 1–830.

Miranda Estrampes, M. (2022). La prueba digital en el proceso penal iberoamericano: Concepto, tratamiento normativo y valoración judicial. Bosch Editor.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2), 1-21. https://doi.org/10.1177/2053951716679679

Mittelstadt, B., Russell, C., & Wachter, S. (2022). Explaining explanations in AI. Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT), 279–288. https://doi.org/10.1145/3287560.3287574

Mokander, J., & Floridi, L. (2021). Ethics-based auditing to develop trustworthy AI. Minds and Machines, 31(2), 323-347. https://doi.org/10.1007/s11023-021-09557-8

Morón Lerma, E. (2022). Internet y derecho penal: hacking y otras conductas ilícitas en la red. Aranzadi-Thomson Reuters, pp. 1–412.

Nieto Martín, A. (2021). Manual de cumplimiento penal en la empresa (2.ª ed.). Tirant lo Blanch.

Nieto Martín, A. (2021). Responsabilidad social, gobierno corporativo y derecho penal de empresa. Atelier Penal, pp. 1–350.

Nussbaum, M. C. (2006). Frontiers of justice: Disability, nationality, species membership. Harvard University Press.

NIST – National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1

OCDE. (2023). OECD Principles on AI (updated). OECD Publishing. https://doi.org/10.1787/9789264304789-en

OCDE. (2023). OECD Framework for the Classification of AI Systems. Organisation for Economic Co-operation and Development. https://doi.org/10.1787/cb6d9eca-en

O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown Publishers.

Ortega, R., & Salinas, P. (2023). Inteligencia artificial y proceso penal en América Latina: desafíos regulatorios y garantistas. Revista Latinoamericana de Derecho Penal, 12(2), 45-78. https://doi.org/10.22197/rldp.v12i2.1423

Ortega, A., & Salinas, P. (2023). Gobernanza de la inteligencia artificial en América Latina: Desafíos regulatorios y perspectivas comparadas. Revista Latinoamericana de Derecho Tecnológico, 4(2), 45–89. https://doi.org/10.35953/rldti.v4i2.215

Palazzi, P. A. (2022). Derecho penal informático (2a ed.). Abeledo Perrot, pp. 1–620.

Palazzi, P. A. (2022). Los delitos informáticos en el Código Penal: Análisis de la Ley 26.388 (3.ª ed. actualizada). Abeledo Perrot.

Palomo, D. (2024). La regulación de la inteligencia artificial en América Latina: una visión comparada. Revista Chilena de Derecho y Tecnología, 13(1), 1–34. https://doi.org/10.5354/0719-2584.2024.73921

Palomo, R. (2024). Marcos regulatorios de inteligencia artificial en América Latina: Estado de la cuestión y propuestas de convergencia. Ius et Scientia, 10(1), 45–82.

Pasquale, F. (2020). New Laws of Robotics: Defending Human Expertise in the Age of AI. Harvard University Press.

Pasquale, F. (2020). The black box society: The secret algorithms that control money and information. Harvard University Press.

Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. Trends in Cognitive Sciences, 25(5), 388–402. https://doi.org/10.1016/j.tics.2021.02.007

Raeini, A. Q., Mostaghim, A., & Kossowska, M. (2023). Forensic AI: Challenges and opportunities in digital evidence processing. Digital Investigation, 44, 301465. https://doi.org/10.1016/j.fsidi.2022.301465

Raeini, A. Q., Mofrad, M. R. K., & Blunt, M. J. (2023). Standards for the admissibility of digital forensic evidence in criminal proceedings: A comparative analysis. Digital Investigation, 44, 101813. https://doi.org/10.1016/j.diin.2022.101813

Raeini, A., Valentin, J., & Mori, G. (2023). Artificial intelligence in digital forensics: Challenges and opportunities for evidence admissibility. Digital Investigation, 44, 301482. https://doi.org/10.1016/j.diin.2023.301482

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 33-44. https://doi.org/10.1145/3351095.3372873

Raso, F., Hilligoss, H., Krishnamurthy, V., Bavitz, C., & Kim, L. (2022). Artificial Intelligence & Human Rights: Opportunities & Risks. Berkman Klein Center Research Publication, 2018-6. https://doi.org/10.2139/ssrn.3259344

Rawls, J. (1971). A theory of justice. Harvard University Press.

Reyes, C., & Lutteral, M. (2021). Cadena de custodia de evidencia digital: Fundamentos técnicos y jurídicos para su implementación en los sistemas de justicia latinoamericanos. Revista Iberoamericana de Ciencias Penales, 8(1), 112–145.

Richardson, R., Schultz, J., & Crawford, K. (2019). Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice. New York University Law Review Online, 94, 192-233.

Rid, T. (2020). Active measures: The secret history of disinformation and political warfare. Farrar, Straus and Giroux, pp. 1–512.

Roxin, C. (2020). Derecho penal: parte general. Tomo I: fundamentos. La estructura de la teoría del delito (2a ed., trad. D. M. Luzón Peña, M. Díaz y García Conlledo & J. de Vicente Remesal). Civitas, pp. 1–1072.

Roxin, C. (2020). Autoría y dominio del hecho en derecho penal (9.ª ed., trad. J. Cuello Contreras & J. Serrano González de Murillo). Marcial Pons, pp. 1–820.

Roxin, C., & Greco, L. (2020). Strafrecht Allgemeiner Teil Band I: Grundlagen, der Aufbau der Verbrechenslehre (5.ª ed.). C.H. Beck.

Sain, G. (2021). El cibercrimen: políticas de prevención y control del delito informático. Editorial Hammurabi, pp. 1–356.

Sain, M. (2021). El crimen en la era digital: Ciberdelito, política criminal e inteligencia artificial en América Latina. Capital Intelectual.

Scherer, M. U. (2021). Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies. Harvard Journal of Law & Technology, 29(2), 353–400. https://doi.org/10.2139/ssrn.2609777

Sieber, U. (2020). Legal order in a world of artificial intelligence. Max Planck Institute for Foreign and International Criminal Law Research Paper Series, No. 2020-01, pp. 1–72.

Sieber, U. (2020). International criminal law in cyberspace: Challenges and perspectives. Max Planck Institute for Foreign and International Criminal Law, pp. 1–312.

Sieber, U. (2020). The emergence of information law: Foundations, aims, and areas of a new legal field. Max Planck Institute for Foreign and International Criminal Law.

Siems, M. (2018). Comparative Law (3.ª ed.). Cambridge University Press, pp. 1–375.

Solove, D. J., & Schwartz, P. M. (2022). Privacy law fundamentals (6th ed.). IAPP Publications, pp. 1–342.

Solum, L. B. (2021). Legal personhood for artificial intelligences. North Carolina Law Review, 70(4), 1231–1287.

Turner, J. (2019). Robot Rules: Regulating Artificial Intelligence. Palgrave Macmillan, pp. 1–356.

UNESCO. (2021). Recommendation on the Ethics of Artificial Intelligence. UNESCO. https://doi.org/10.54677/MNMH8546

UNESCO. (2021). sobre la ética de la inteligencia artificial. Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura. https://unesdoc.unesco.org/ark:/48223/pf0000381137_spa

UNODC. (2023). Cybercrime module series: E4J University module series on cybercrime. United Nations Office on Drugs and Crime. https://www.unodc.org/e4j/en/cybercrime/index.html

Vladeck, D. C. (2022). Machines without principals: Liability rules and artificial intelligence. Washington Law Review, 89(1), 117–150.

Wachter, S., Mittelstadt, B., & Russell, C. (2020). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), 841–887. https://doi.org/10.2139/ssrn.3063289

World Economic Forum. (2024). Global cybersecurity outlook 2024. World Economic Forum, pp. 1–56. https://www.weforum.org/publications/global-cybersecurity-outlook-2024/

Zaffaroni, E. R. (2019). El enemigo en el derecho penal (2ª ed.). EDIAR.

Zaffaroni, E. R. (2019). Manual de derecho penal: Parte general (3.ª ed.). Ediar.

Završnik, A. (2021). Algorithmic justice: Algorithms and big data in criminal justice settings. European Journal of Criminology, 18(5), 623-642. https://doi.org/10.1177/1477370819876762

Završnik, A. (2021). Criminal justice, artificial intelligence systems, and human rights. ERA Forum, 20(4), 567–583. https://doi.org/10.1007/s12027-020-00602-0

Zittrain, J. (2019). The hidden costs of automated thinking. The New Yorker, July 23. https://www.newyorker.com/tech/annals-of-technology/the-hidden-costs-of-automated-thinking

Zuboff, S. (2021). The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. PublicAffairs.

Zweigert, K., & Kötz, H. (2020). Introduction to Comparative Law (3.ª ed., trad. T. Weir). Oxford University Press, pp. 1–714.

Septiembre 2026 - CID - Centro de Investigación y Desarrollo
Copyright © - CID - Centro de Investigación y Desarrollo
Copyright del texto © 2026 de Autores
Formato: PDF
Tamaño: A4 210 x 297 mm
Requisitos de sistema: Adobe Acrobat Reader
Modo de acceso: World Wide Web
biblioteca.ciencialatina.org
editorial@ciencialatina.org
Atención por WhatsApp al +52 1 55 9435 7573