

Inteligencia artificial y derecho penal

Fernando Velásquez V.*

Sumario. I. Introducción. II. Punto de partida para el análisis: el marco conceptual. III. El impacto en las construcciones del derecho penal: *a. Aspectos generales, b Otras manifestaciones, c. Conclusión, d. La reconstrucción objetiva de los hechos.* IV. La IA en el proceso penal. V. La IA en la investigación criminal: *a. Herramientas de reconocimiento facial; b. El análisis predictivo; c. La evidencia digital.* VI. Conclusiones. VII Referencias.

Resumen. Este trabajo examina el impacto transformador de la inteligencia artificial (IA) en el ámbito del derecho penal y procesal penal, tanto desde una perspectiva teórica como práctica; se analizan las aplicaciones actuales de la IA en la investigación criminal, la predicción de delitos, la evaluación de riesgos, la toma de decisiones judiciales y la ejecución penal. El texto aborda los desafíos éticos, jurídicos y constitucionales que surgen de la introducción de estas tecnologías, incluyendo cuestiones de transparencia, sesgo algorítmico, privacidad y debido proceso. Se examina el impacto de estos sistemas en los principios fundamentales del derecho penal, como la presunción de inocencia, la proporcionalidad y la individualización de la pena. A través de un análisis comparativo internacional, se identifican las mejores prácticas y se proponen marcos regulatorios para garantizar que la integración de la IA en la justicia penal respete los derechos humanos y los principios del Estado de Derecho.

Abstract. This work examines the transformative impact of artificial intelligence (AI) in the field of criminal law and criminal procedure, from both theoretical and practical perspectives; it analyzes the current applications of AI in criminal investigation, crime prediction, risk assessment, judicial decision-making, and criminal enforcement. The text addresses the ethical, legal, and constitutional challenges that arise from the introduction of these technologies, including issues of transparency, algorithmic bias, privacy, and due process. It examines the impact of these systems on the fundamental principles of criminal law, such as the presumption of innocence, proportionality, and individualization of punishment. Through a comparative international analysis, best practices are identified and regulatory

* Asesor Académico Universidad Sergio Arboleda, Bogotá, Colombia.

frameworks are proposed to ensure that the integration of AI in criminal justice respects human rights and the principles of the Rule of Law.

Palabras clave. Inteligencia artificial, derecho penal, proceso penal, justicia predictiva, ética algorítmica, administración de justicia, derechos fundamentales.

Keywords. Artificial intelligence, criminal law, criminal procedure, predictive justice, algorithmic ethics, administration of justice, fundamental rights.

I. INTRODUCCIÓN

La intersección entre la inteligencia artificial (IA) y el sistema de justicia penal representa uno de los desarrollos más significativos y controvertidos en la evolución del derecho contemporáneo; este trabajo explora las profundas transformaciones que la integración de la IA genera en el derecho penal y procesal penal, tanto en su dimensión teórica como en sus aplicaciones prácticas dentro de la administración de justicia criminal.

Y no es para menos: La revolución digital que caracteriza al siglo XXI ha penetrado en prácticamente todos los ámbitos de la vida social, y el sistema jurídico-penal no ha quedado al margen de esta transformación; desde luego, la introducción de sistemas de IA en la justicia penal promete mejorar la eficiencia, precisión y objetividad en la toma de decisiones, optimizar la asignación de recursos escasos, y, potencialmente, reducir los sesgos humanos que afectan a la administración de justicia. Sin embargo, estas promesas vienen acompañadas de riesgos significativos y desafíos fundamentales para los principios que sustentan el derecho penal moderno.

La tecnología no es neutral; los sistemas de IA incorporan valores, prioridades y sesgos que reflejan tanto las decisiones conscientes de sus diseñadores como las limitaciones inherentes a los datos con los que son entrenados. En el contexto del derecho penal, donde están en juego los derechos fundamentales y la libertad de las personas, estas cuestiones adquieren una relevancia crítica que trasciende el mero debate técnico, para convertirse en un asunto de profunda importancia jurídica, ética y política.

Ciertamente la incorporación de la IA en el sistema penal está ocurriendo a un ritmo acelerado, a menudo sin un marco normativo adecuado ni una reflexión suficiente sobre sus implicaciones, por lo cual escenarios como este encuentro académico son muy importantes para jalonar el debate. Sistemas de predicción de riesgos, herramientas de vigilancia algorítmica, *software* de reconocimiento facial y programas de apoyo a las decisiones judiciales son ahora utilizados en diversas jurisdicciones, lo que genera transformaciones silenciosas pero profundas en la forma como se investigan los delitos, se procesa a los acusados y se imponen y ejecutan las sanciones penales.

En cualquier caso, debe quedar claro que el uso de IA en el derecho penal tiene potencial, pero también plantea riesgos significativos; es esencial, pues, un marco legal adecuado para garantizar que su aplicación sea justa, ética y respetuosa con los derechos humanos (Staffler & Jany, 2020). El derecho –y con él el derecho penal– y la sociedad debe evolucionar para abordar los desafíos planteados por sistemas de IA cada vez más sofisticados (Rademacher, 2025).

Este ensayo crítico aborda estas cuestiones desde una perspectiva interdisciplinaria que combina el análisis jurídico-dogmático con consideraciones técnicas, éticas y sociales; por eso, se analizan tanto los beneficios potenciales como los riesgos y desafíos que plantea la integración de la IA en el sistema penal, y se proponen vías para un empleo responsable de la misma que preserve los valores fundamentales del Estado de Derecho. En última instancia, el objetivo de esta reflexión es tratar de contribuir a un debate informado sobre cómo aprovechar el potencial transformador de la IA para mejorar la justicia penal, pero sin sacrificar los principios de dignidad humana, igualdad, acto, culpabilidad, transparencia y debido proceso, que son la base de un sistema penal legítimo en una sociedad democrática, que aboga por un programa penal de la constitución respetuoso de los derechos de la persona humana.

II. PUNTO DE PARTIDA PARA EL ANÁLISIS: EL MARCO CONCEPTUAL

La inteligencia artificial (IA) es uno de los avances tecnológicos más impactantes de nuestra era. Hoy, más que vivir en una ‘sociedad del riesgo’, como propuso Beck en 1985, estamos en una ‘sociedad de la innovación’ (Ibold, 2024) que no solo experimenta cambios, sino que los impulsa activamente en áreas como la economía, lo social y el derecho. Por ello, antes de analizar sus implicaciones específicas en el derecho penal, resulta fundamental establecer un marco conceptual claro y preciso sobre lo que se entiende por IA y cómo esta herramienta ha evolucionado hasta convertirse en un instrumento llamado a generar una transformación potencial de todo el sistema jurídico.

La definición clásica propuesta por John McCarthy, considerado uno de los padres fundadores de la IA, la describe como “la ciencia e ingeniería de hacer máquinas inteligentes, especialmente programas informáticos inteligentes” (McCarthy, 2007, p. 2). Esta noción, aunque útil como punto de partida, resulta insuficiente para capturar la complejidad y diversidad de lo que hoy se entiende por tal. De esta forma, un acercamiento al asunto más contemporáneo y relevante para el ámbito jurídico entiende que la IA se refiere a sistemas computacionales capaces de realizar tareas que normalmente requieren inteligencia humana, como la percepción visual, el reconocimiento del habla, la toma de decisiones, la traducción de idiomas y el aprendizaje automático a partir de datos (Russell & Norvig, 2008; 2021).

Sin embargo, la evolución histórica de la IA puede dividirse en varias etapas fundamentales que ayudan a comprender su estado actual y posibles trayectorias futuras. El momento fundacional, entre 1950 y 1956, estableció las bases conceptuales de la IA con los trabajos de Alan Turing, particularmente su artículo *"Computing Machinery and Intelligence"* (1950); incluso este período dio lugar a trabajos literarios como el de Giovanni Papini, en su *Libro Negro* (1973), con su conocido relato del "Tribunal electrónico" que data de 1952. El término "inteligencia artificial" fue acuñado formalmente en la conferencia de Dartmouth de 1956, organizada por McCarthy, Minsky, Rochester y Shannon. Posteriormente, entre 1956 y 1974, se produjo una primera ola de entusiasmo caracterizada por avances significativos en programas que resolvían problemas algebraicos, demostraban teoremas y jugaban ajedrez. Los sistemas expertos basados en reglas comenzaron a desarrollarse, mostrando capacidad para emular el razonamiento humano en dominios específicos.

Entre 1974 y 1980 se experimentó el primer "invierno de la IA", cuando las limitaciones técnicas y la sobrevaloración de las capacidades de la IA condujeron a una reducción del interés y la financiación de la investigación en este campo. Sin embargo, durante la década de 1980 hasta 1987, se produjo el auge de los sistemas expertos, oportunidad en la cual la IA comercial floreció con el desarrollo de sistemas que incorporaban conocimiento especializado para resolver problemas específicos, incluyendo algunas aplicaciones en el ámbito legal. Un segundo "invierno de la IA" ocurrió entre 1987 y 1993, cuando de nuevo las expectativas excesivas llevaron a la desilusión y la reducción de inversiones.

Entre 1993 y 2011, la IA se reorientó hacia métodos estadísticos y probabilísticos, sentando las bases para el aprendizaje automático moderno; de este momento –una verdadera primavera– data el célebre enfrentamiento entre el campeón mundial de ajedrez Gari Kasparov y el computador *Deep Blue* diseñado por la IBM, que –hacia el mes de febrero de 1996– condujo a la derrota de la máquina por 4 a 2, tras seis reñidos encuentros; la felicidad duró poco porque en mayo de 1997 una versión mejorada del ordenador derrotó al campeón, en una nueva disputa (Campbell *et al.*, 2002).

Finalmente, desde 2012 hasta el presente se observa una revolución del aprendizaje profundo, caracterizada por el desarrollo de arquitecturas neuronales de múltiples capas, combinado con el incremento exponencial en la capacidad computacional y la disponibilidad de datos masivos. Esta fase ha propulsado avances sin precedentes en el reconocimiento de imágenes, el procesamiento del lenguaje natural, los sistemas de recomendación (esto es, algoritmos y técnicas de IA diseñados para predecir y sugerir elementos –productos, servicios, contenidos o información– que podrían ser de interés para un usuario específico) y otros dominios, permitiendo la integración de la IA en el sistema jurídico-penal de formas previamente inimaginables.

En el contexto específico del derecho penal, la IA ha evolucionado desde simples sistemas expertos que codificaban normativas penales hacia sofisticados algoritmos predictivos que analizan patrones de criminalidad, evalúan riesgos procesales y asis-

ten en la toma de decisiones judiciales. Esta evolución no representa solo un cambio cuantitativo en la capacidad computacional, sino una transformación cualitativa en la forma en que la tecnología interactúa con el sistema de justicia penal, planteando nuevos desafíos para los principios fundamentales del derecho penal como la presunción de inocencia, la proporcionalidad y la transparencia procesal, etc.

Asimismo, en el ámbito jurídico-penal resulta esencial distinguir entre diferentes categorías de sistemas de IA, porque sus capacidades, limitaciones e implicaciones varían significativamente (Martínez, 2019). La mayoría de los existentes hoy en el ámbito penal pertenecen, de un lado, a la categoría de IA Estrecha o Débil (*Narrow or Weak AI*), diseñados para realizar tareas específicas sin conciencia genuina ni comprensión semántica profunda. Estos modelos incluyen sistemas basados en reglas que codifican el conocimiento jurídico explícito mediante reglas lógicas *if-then* (si entonces...) utilizados para verificar el cumplimiento normativo y asistir en la calificación jurídica de hechos; su principal ventaja radica en su transparencia y explicabilidad, aunque su rigidez les impide adaptarse a la complejidad y ambigüedad inherentes al razonamiento jurídico-penal.

También, los sistemas de aprendizaje automático (*machine learning*) procesan grandes volúmenes de datos históricos para detectar patrones; en el ámbito penal, ellos se usan para predecir la probabilidad de que una persona reincida en el delito, evaluar riesgos durante un proceso judicial o detectar patrones de comportamiento criminal. Incluso los sistemas de procesamiento de lenguaje natural (NLP) analizan y generan lenguaje humano, facilitando el análisis automatizado de documentos jurídicos, jurisprudencia y testimonios. Por su parte, los sistemas de visión computacional interpretan y analizan información visual, incluyendo tecnologías de reconocimiento facial, análisis de escenas del crimen y la verificación de evidencia audiovisual.

De otro lado, la llamada IA General o Fuerte (*General or Strong AI*) representaría sistemas hipotéticos con capacidades cognitivas comparables a las humanas, incluyendo la comprensión contextual, el razonamiento abstracto y la transferencia de conocimiento entre dominios; por supuesto, aunque estos modelos todavía no existen, su eventual desarrollo plantearía cuestiones profundas sobre la responsabilidad penal, la agencia moral (entiéndase, en un plano filosófico, la capacidad de discernir entre el bien y el mal, y actuar en consecuencia con cierto grado de libertad y comprensión de las normas morales, por parte de los nuevos entes) y la posible subjetividad jurídica de entidades artificiales.

En ese contexto, según su funcionamiento y transparencia, se pueden clasificar los sistemas en “cajas negras”, cuyo funcionamiento interno no es transparente ni fácilmente interpretable, como ocurre con muchos algoritmos de aprendizaje profundo; y en “cajas blancas”, que permiten comprender y explicar el proceso mediante el cual llegan a una conclusión determinada. Los primeros modelos plantean serios desafíos para los principios fundamentales del derecho penal como son los atinentes a la motivación de las decisiones judiciales y el derecho a la defensa, mientras que

los segundos son preferibles desde la perspectiva del debido proceso y la seguridad jurídica.

Según su autonomía decisoria, se distingue entre “sistemas de apoyo a la decisión”, que proporcionan información y recomendaciones pero la decisión final recae en un juzgador humano, y “sistemas de decisión automatizada”, que toman decisiones sin intervención humana directa. La mayoría de las aplicaciones actuales en el sistema penal corresponden a la primera categoría, mientras que la utilización de sistemas de decisión completamente automatizada en el ámbito penal resulta altamente controvertida, especialmente cuando ellos afectan derechos fundamentales como la libertad personal.

Asimismo, para comprender adecuadamente el impacto de la IA en el derecho penal, resulta fundamental conocer los principios técnicos básicos que subyacen a los sistemas más utilizados actualmente en el ámbito jurídico. De hecho, el *machine learning* (aprendizaje automático) y su subconjunto más avanzado, el *deep learning* (aprendizaje profundo), constituyen las tecnologías predominantes en las aplicaciones penales contemporáneas.

El *machine learning* representa un paradigma fundamentalmente distinto a la programación tradicional (Forero-Corba & Negre Bannasar, 2024). Por eso, mientras que en esta última los desarrolladores codifican explícitamente reglas para resolver problemas específicos, en el aprendizaje automático los algoritmos “aprenden” patrones a partir de datos históricos para realizar predicciones o tomar decisiones sobre nuevos casos. Este enfoque resulta muy valioso en dominios donde las reglas son difíciles de formalizar o existen patrones subyacentes demasiado complejos para ser codificados manualmente, características que definen muchos aspectos del derecho penal.

Los principales paradigmas del *machine learning* relevantes para aplicaciones jurídico-penales incluyen el aprendizaje supervisado, en cuya virtud el algoritmo aprende a partir de ejemplos etiquetados, estableciendo correlaciones entre características (*inputs*) y resultados (*outputs*). De esta forma, en el contexto penal podría entrenarse un sistema con datos históricos de casos resueltos para predecir, por ejemplo, el riesgo de reincidencia o la probabilidad de fuga procesal de un imputado; incluso, los métodos más utilizados incluyen la regresión logística, los árboles de decisión y el algoritmo *Random Forests* (bosque aleatorio) y las máquinas de vectores de soporte (SVM).

El aprendizaje no supervisado permite que el algoritmo identifique patrones y estructuras latentes en datos no etiquetados; en la investigación criminal, estas técnicas pueden detectar anomalías en transacciones financieras o identificar patrones delictivos previamente desconocidos. Los métodos principales incluyen el *clustering* (agrupamiento) que junta casos similares, y revela potenciales tipologías delictivas o patrones criminológicos, amén de la detección de anomalías que identifican des-

viaciones significativas de patrones normales existentes, que es útil para detectar fraudes o comportamientos sospechosos.

Por su parte, el aprendizaje por refuerzo permite que el algoritmo aprenda mediante ensayo y error, recibiendo “recompensas” por acciones deseables; este paradigma tiene aplicaciones potenciales en la optimización de estrategias de vigilancia y asignación de recursos policiales.

El *deep learning* (aprendizaje profundo) ha revolucionado campos como el reconocimiento facial, el procesamiento del lenguaje natural y el análisis de imágenes, todos con aplicaciones directas en el sistema penal. Incluso, las redes neuronales profundas imitan estructuralmente el funcionamiento del cerebro humano mediante capas interconectadas de “neuronas” artificiales, permitiendo el modelado de relaciones extremadamente complejas y no lineales.

También, arquitecturas de *deep learning* particularmente relevantes para aplicaciones penales incluyen las redes neuronales convolucionales (CNN), especializadas en procesar datos con estructura de cuadrícula, como imágenes, fundamentales para las tecnologías de reconocimiento facial, análisis automático de evidencia visual y sistemas de vigilancia. Asimismo, las redes neuronales recurrentes (RNN) y los *Transformers* (transformadores), diseñados para procesar datos secuenciales como texto o habla, esenciales para analizar documentos jurídicos, transcripciones de interrogatorios y comunicaciones potencialmente delictivas.

Aquí se debe destacar que los modelos de lenguaje de gran escala, como GPT (*Generative Pre-trained Transformer*) y BERT (*Bidirectional Encoder Representations from Transformers*), han revolucionado la comprensión y generación del lenguaje natural, al permitir análisis sofisticados de documentación jurídica, extracción de información relevante de grandes *corpus* legales y la asistencia en la redacción de documentos jurídicos.

Estos sistemas presentan características técnicas que plantean desafíos específicos para el derecho penal (Bell *et al.*, 2021); la opacidad de los algoritmos, especialmente en modelos de aprendizaje profundo, se debe a su complejidad intrínseca, lo que dificulta entender cómo se toman las decisiones, incluso para los desarrolladores. Estos sistemas funcionan como ‘cajas negras’, donde millones de cálculos interactúan de maneras que son casi imposibles de interpretar para los humanos; esta característica técnica choca frontalmente con principios como la transparencia judicial y el derecho a conocer los fundamentos de las decisiones que afectan derechos fundamentales.

Además, la dependencia de datos históricos implica que los sistemas de *machine learning* aprenden de ellos, potencialmente perpetuando y amplificando sesgos preexistentes en el sistema penal; los sesgos en los sistemas de IA a menudo se derivan de los datos de entrenamiento, que pueden reflejar desigualdades históricas o sociales, perpetuando así discriminaciones existentes. Desde luego, si los datos

de entrenamiento reflejan disparidades raciales, socioeconómicas o geográficas en la aplicación de la ley, el algoritmo resultante incorporará estos sesgos de forma aparentemente objetiva y científica.

Otra limitación importante es la confusión entre *correlación* y *causalidad*, porque estos sistemas identifican correlaciones estadísticas sin comprender necesariamente relaciones causales, lo que puede conducir a conclusiones espurias cuando se aplican a fenómenos sociales complejos como la criminalidad. Es crucial recordar que los sistemas de IA identifican correlaciones estadísticas, pero no necesariamente relaciones causales, lo que puede llevar a conclusiones erróneas en contextos sociales complejos. Además, los modelos de IA funcionan mejor en condiciones similares a sus datos de entrenamiento debido a limitaciones en la generalización, lo cual significa que pueden fallar en situaciones nuevas o atípicas, característica que los hace problemáticos en relación con un sistema penal que debe garantizar la seguridad jurídica y el trato equitativo.

En fin, comprender estos fundamentos técnicos resulta indispensable para evaluar críticamente el potencial, las limitaciones y los riesgos de la introducción de sistemas de IA en contextos jurídico-penales, amén de diseñar salvaguardas adecuadas que preserven los principios fundamentales del Estado de Derecho en la era digital. El anterior es, pues, el marco conceptual básico que permite abordar la problemática de esta ponencia; ahora mírense los influjos en el derecho penal sustantivo, procesal y en la investigación penal.

III. EL IMPACTO EN LAS CONSTRUCCIONES DEL DERECHO PENAL

a. Aspectos generales

Como construcción teórica que busca sistematizar, interpretar y aplicar las normas penales, la dogmática penal se enfrenta al desafío de incorporar y responder a nuevas realidades que la IA introduce en el panorama jurídico (Barrio Andrés, 2019). Por eso, se debe analizar cómo la IA incide en las construcciones teóricas fundamentales del derecho penal, parte general, y proponer vías para una adaptación constructiva de esta disciplina a la nueva realidad tecnológica; desde luego, el sacudón que hoy se vive también abarca la parte especial del mismo, en atención a que muchas de sus figuras punibles se van quedando obsoletas, como lo ejemplifican los casos del fraude y la extorsión (Glaser, 2024).

Así las cosas, se parte de la premisa de que la IA no solo es un objeto de regulación en el derecho penal, sino que también transforma sus categorías conceptuales y metodológicas fundamentales; por eso, se deben mirar las transformaciones de sus conceptos y elaboraciones tradicionales como los atinentes a las fuentes del derecho penal, la conducta, el nexo de causalidad y la imputación objetiva, el bien jurídico,

la tipicidad, la antijuridicidad, la autoría, la culpabilidad, la pena, etc. (Velásquez, 2025, p. 303). Por ello, se hace necesario reflexionar sobre el futuro de la dogmática penal en un contexto de creciente integración tecnológica así sea para delinear los nuevos marcos teóricos que permitan adaptar el derecho penal a esta realidad emergente.

De este forma, lo primero a tener en cuenta es la transformación del objeto de estudio de la ciencia del derecho penal, entendida como la disciplina que investiga de forma sistemática el conjunto de normas jurídico-penales y las categorías conceptuales que permiten su aplicación. Como se sabe, las construcciones doctrinales tradicionales parten de un esquema antropocéntrico para el cual el ser humano, como agente moral autónomo, constituye el único sujeto capaz de cometer delitos y ser penalmente responsable (Muñoz Conde & García Arán, 2022); sin embargo, con la introducción de la IA ella experimenta una transformación significativa de su objeto.

También el método jurídico-penal afronta una transformación significativa, dado que la interpretación teleológica, la construcción sistemática y la argumentación jurídica –pilares metodológicos de la dogmática penal contemporánea– requieren incorporar herramientas analíticas propias de las ciencias computacionales, la filosofía de la tecnología y la ética de la IA. Al respecto, es importante señalar que las fuentes tradicionales del conocimiento penal se ven ahora complementadas por otras que emergen del ecosistema digital: algoritmos de análisis jurisprudencial, sistemas predictivos de riesgo criminal, y herramientas de análisis de evidencia digital, entre otros. Estos instrumentos no solo facilitan la labor del jurista, sino que transforman cualitativamente su aproximación metodológica al fenómeno criminal.

En este contexto, el principio de legalidad penal (*nullum crimen, nulla poena sine lege*), pilar fundamental del derecho penal moderno, experimenta una renovación ante la influencia de la IA; es más: la creciente complejidad técnica de los sistemas de IA plantea desafíos significativos para la taxatividad y accesibilidad de las normas penales. Piénsese, por ejemplo, en lo que sucede en materia de normas penales en blanco en relación con los complementos del supuesto de hecho con la la remisión directa o indirecta de ámbitos completos de la legislación penal a lo que determinen reglas técnicas, o bien a estándares éticos, elaborados por asociaciones profesionales u otras entidades privadas (Silva Sánchez, 2025, p. 342).

La IA también incide en las fuentes materiales del derecho penal, entendidas como los factores que influyen en el contenido de las normas penales; los estándares técnicos de seguridad informática, los principios éticos para el desarrollo de IA y las directrices de gobernanza algorítmica adquieren relevancia como referentes normativos que informan la creación e interpretación de los tipos penales. Se produce, así, una hibridación normativa donde las fuentes tradicionales (legislación, jurisprudencia, costumbre) interactúan con estándares técnicos y principios de ética computacional, dando lugar a un sistema de fuentes más complejo y multinivel.

Desde luego, uno de los impactos más profundos de la IA en la teoría penal concierne al ámbito de validez personal de la ley penal, es decir, a la determinación de quiénes pueden ser sujetos de imputación jurídico-penal, un ámbito tradicionalmente circunscrito a personas físicas y, en determinados ordenamientos, a personas jurídicas. Por ello, en ese contexto, la IA plantea interrogantes fundamentales: ¿Puede un sistema de IA ser considerado sujeto de imputación penal, máxime si carece de conciencia y agencia moral, lo cual hace que la atribución de responsabilidad penal directa sea un desafío filosófico y jurídico? ¿Debe ampliarse el ámbito de validez personal de la norma penal para incluir a los “agentes artificiales”? A no dudarlo, la discusión está abierta y, por ahora, se deberá rechazar esa posibilidad.

Como es obvio, ante la dificultad de considerar a los sistemas de IA como sujetos penales autónomos emerge la necesidad de desarrollar modelos de imputación basados en la *responsabilidad distribuida*, que atribuyan responsabilidad penal a los diversos actores humanos involucrados en el diseño, programación, entrenamiento, implementación y supervisión de sistemas de IA. Esta perspectiva implica una expansión horizontal del ámbito de validez personal, no mediante la inclusión de nuevos “sujetos artificiales”, sino a través de la ampliación de la esfera de responsabilidad humana en contextos tecnológicamente mediados.

Incluso, para aludir no a las personas naturales sino a las empresas se puede discutir el asunto en el seno de una responsabilidad penal por el producto –que aquí es la IA– y precisar si su sede es el derecho penal o un diseño intermedio entre el derecho penal y el derecho administrativo sancionador que, en países como Alemania, está contenido en la llamada Ley de Contravenciones al Orden (Ordnungswidrigkeitengesetz -OWiG- promulgada el 24 de mayo de 1968 en Alemania Occidental y que entró en vigor el 1 de enero de 1969) (véase, Ibold, 2024).

En cuanto se refiere a la teoría del delito, como constructo dogmático que sistematiza los elementos constitutivos del hecho punible, también se experimenta una profunda renovación ante el fenómeno de la IA. El concepto de acción, primer escalón de la estructura del delito, requiere replanteamientos significativos y, está claro por ahora, dentro de los sujetos o entes que no realizan acciones con relevancia jurídico penal será necesario incluir “los procesos que surgen de máquinas provistas de inteligencia artificial –de procesamiento artificial de macrodatos–. Este es el caso, por ejemplo, de los coches autopilotados o de los sistemas de armamento inteligente; pero también el de los procesadores que establecen los precios de las mercancías. En ellos no existe una teleología propia, de modo que, aun en el caso de que sean razonadores teóricos, carecen de razón práctica –capacidad de acción y de autodeterminación conforme a sentido” (Silva Sánchez, 2025, p. 656; Velásquez, 2025, p. 337).

Como se recordará, las concepciones tradicionales de la acción penal (causal, final, social, personal, negativa) parten de la premisa de un comportamiento humano voluntario externamente perceptible. Sin embargo, las decisiones algorítmicas

pueden generar efectos jurídicos relevantes sin que exista propiamente una “acción humana” en sentido clásico; de allí, emerge la necesidad de desarrollar en un futuro próximo un concepto de acción penal que comprenda las diversas modalidades de interacción humano-máquina, desde la programación directa hasta la supervisión de sistemas autónomos de aprendizaje.

También la tipicidad y con ella el tipo penal, como elemento configurador del delito que describe la conducta mandada o prohibida, afronta el reto de captar adecuadamente acciones y omisiones en entornos computacionales complejos; hoy los tipos penales tradicionales, diseñados para conductas humanas directas, resultan frecuentemente inadecuados para escenarios donde la causalidad es algorítmica y distribuida. Se produce así una expansión de los tipos de peligro abstracto y de los delitos de infracción de deber, como estrategia para abordar riesgos sistémicos derivados del uso de IA; piénsese, no más, lo que sucede en materia de delitos informáticos (Posada Maya, 2017). Este fenómeno, sin embargo, genera tensiones con principios tradicionales como los de lesividad y *ultima ratio*.

Ahora bien, ubicados en este contexto debe señalarse que la creación de nuevas figuras punibles en la ley penal, llamadas a castigar diversas conductas que se realizan en el escenario propio de la IA, obliga a redefinir el papel de los bienes jurídicos, sobre todo los colectivos (Vallas Prieto, 2023). La teoría del bien jurídico, concepto tradicionalmente concebido como interés o valor esencial para la coexistencia social y el desarrollo de la personalidad individual, requiere una reinterpretación en al menos tres dimensiones fundamentales:

En primer lugar, emerge la necesidad de identificar y conceptualizar nuevos bienes jurídicos específicamente vinculados al entorno digital y algorítmico. La integridad de los sistemas informáticos, la autodeterminación informativa, la veracidad algorítmica o la transparencia en la toma de decisiones automatizadas constituyen valores emergentes que podrían requerir tutela penal autónoma, trascendiendo la mera reconceptualización de bienes jurídicos tradicionales.

En segundo término, los bienes jurídicos clásicos experimentan una transformación cualitativa en cuanto a sus formas de vulneración. La vida, la integridad física, la intimidad o el patrimonio, pueden verse afectados mediante nuevas modalidades de agresión mediadas por sistemas inteligentes, lo que exige reconsiderar el alcance de su protección y los criterios de determinación de su lesividad. Finalmente, la abstracción creciente de ciertos bienes jurídicos supraindividuales –como la seguridad colectiva o la confianza en el funcionamiento del sistema socioeconómico– se intensifica en el contexto de la IA, planteando interrogantes sobre la legitimidad y efectividad de su protección penal anticipada mediante delitos de peligro abstracto, especialmente considerando la complejidad predictiva de los sistemas autónomos.

De igual forma, categorías como la causalidad y la imputación objetiva también requieren replanteamientos. Históricamente, recuérdese, la teoría jurídico-penal se

ha basado en un paradigma de causalidad natural, secuencial y directamente atribuible a un agente humano, incluso cuando se piensa en la forma de concebir la categoría de la imputación objetiva; sin embargo, la incorporación de sistemas de inteligencia artificial introduce una ruptura epistemológica en este modelo. Así, la causalidad, en el contexto de la IA, se manifiesta como un fenómeno reticular, probabilístico y distribuido entre múltiples nodos algorítmicos, ninguno de los cuales puede considerarse aisladamente como condición determinante del resultado.

Esta renovación de la causalidad natural repercute inevitablemente en la construcción normativa de la imputación objetiva. La determinación del riesgo permitido, elemento central de esta teoría, debe reconfigurarse para considerar la gradualidad de la autonomía decisional de estos sistemas, su interacción contextual compleja y el nivel de transparencia algorítmica que presentan; incluso hasta llegar a una *responsabilidad por la innovación* que debería quedar en manos del derecho administrativo sancionador (mírese el debate que propone Ibold, 2024). El principio de confianza, asimismo, requiere una modulación específica que considere las particulares características de delegación decisional en sistemas artificiales.

Adicional a ello, la estructura temporal y multiactoral del desarrollo e implementación de sistemas de IA genera una responsabilidad difusa que desafía los modelos tradicionales de imputación individual; por ejemplo, la intervención secuencial de diseñadores, desarrolladores, entrenadores e implementadores en distintos momentos de la cadena causal complica la determinación precisa de responsabilidades individuales bajo los esquemas clásicos. Ante estos desafíos, resulta imperativo el desarrollo de nuevos modelos de imputación que, sin abandonar los principios fundamentales del derecho penal, se adapten a esta realidad tecnológica emergente; un enfoque prometedor podría orientarse hacia la construcción de criterios de imputación basados en el control efectivo sobre distintos aspectos del sistema (diseño, entrenamiento, implementación) y el establecimiento de deberes cualificados de diligencia específicamente adaptados a las particularidades de los sistemas autónomos.

Asimismo, para aludir al concepto de dolo (*Vorsatz*) –entendido en nuestra normativa como conocer y querer la realización de la conducta típica–, hoy también debe ser reformulado de tal manera que se puedan sustituir los criterios subjetivos por pautas objetivas de previsibilidad y asunción de riesgos, permitiendo su aplicación tanto a humanos como a sistemas artificiales, sin abandonar la estructura dogmática del delito. Otro tanto sucede con el concepto de culpa o imprudencia (*Fahrlässigkeit*), que requerirá de una transformación en cuya virtud se debe transitar hacia modelos más normativos basados en la gobernanza algorítmica, la gestión de incertidumbre y los estándares técnicos validados, desplazando el énfasis desde la previsibilidad de resultados específicos hacia la implementación de procesos adecuados de control, transparencia y supervisión que permitan mantener sistemas complejos dentro de parámetros aceptables de riesgo.

En cuanto a la antijuridicidad, como juicio de desvalor objetivo sobre la conducta típica, también experimenta cambios significativos; por ejemplo, las colisiones de deberes en contextos de decisiones algorítmicas –como los dilemas éticos de los vehículos autónomos– plantean nuevos escenarios para las causas de justificación tradicionales. El estado de necesidad justificante, por ejemplo, requiere adaptaciones cuando es un algoritmo el que debe “decidir” entre bienes jurídicos en peligro; por eso, la programación de algoritmos para situaciones dilemáticas implica una ponderación *ex ante* que difiere sustancialmente de la valoración *ex post* propia del juicio de antijuridicidad tradicional.

Asimismo, la culpabilidad, concebida tradicionalmente como reproche personal o un juicio de exigibilidad basado en la capacidad de comprensión de la ilicitud y de actuación conforme a dicha comprensión, constituye quizás el elemento de la teoría del delito más profundamente desafiado por la IA. Los sistemas de IA carecen de conciencia moral y autodeterminación en sentido humano, lo cual imposibilita un juicio de culpabilidad en sentido tradicional; sin embargo, su creciente autonomía decisiva y su capacidad de aprendizaje generan interrogantes sobre la posibilidad de desarrollar constructos análogos a la culpabilidad para determinados sistemas artificiales. No obstante, en el corto y medio plazo, el impacto principal de la IA en la teoría de la culpabilidad no concierne tanto a la responsabilidad de los sistemas artificiales, sino a la transformación de los criterios de imputación subjetiva aplicables a humanos en contextos tecnológicamente mediados.

Otro aspecto que merece toda la atención es el atinente a la teoría de las consecuencias jurídicas del delito, porque todo indica que se debe producir una reconfiguración de los fines de la pena, cuando se piensa en la justificación, la determinación y la ejecución de penas y medidas de seguridad, lo cual experimenta una reconfiguración sustancial ante el fenómeno de la IA. Las teorías tradicionales sobre los fines de la pena (retribución, prevención general, prevención especial, etc.) requieren adaptaciones cuando el sujeto pasivo de la sanción ha actuado en entornos algorítmicamente mediados; de allí que en la era digital la pena no solo debe dirigirse a potenciales infractores individuales, sino también a corporaciones tecnológicas y diseñadores de sistemas, ampliando así el círculo de destinatarios del mensaje disuasorio penal.

En cuanto a la determinación y ejecución de penas, la IA no solo transforma la fundamentación teórica de las consecuencias jurídicas, sino que también se integra como herramienta en su determinación y ejecución; los sistemas algorítmicos de evaluación de riesgo criminal, cada vez más presentes en jurisdicciones de todo el mundo, inciden directamente en decisiones sobre la libertad condicional, la prisión preventiva y la individualización judicial de la pena. Así, pues, el instrumento de la inteligencia artificial puede ayudar de forma significativa en el proceso de determinación cuantitativa de la pena sin pretender que los algoritmos desplacen al juzgador, esto es, la delegación de juicios valorativos a sistemas algorítmicos en el ámbito

de la determinación de la pena, puede poner en tela de juicio la función garantista del juez como intérprete último del ordenamiento jurídico; de esta manera esos instrumentos, si se usan bien, pueden ayudar a reducir la arbitrariedad y el abuso (Velásquez, 2025, p. 770).

A este respecto, debe señalarse que IA también incide en el catálogo de consecuencias jurídicas disponibles, propiciando el desarrollo de nuevas modalidades sancionadoras adaptadas al entorno digital. Sanciones como la “desconexión digital”, la “reprogramación supervisada” o la “auditoría algorítmica obligatoria” emergen como alternativas a penas tradicionales para determinados delitos relacionados con sistemas informáticos. En este ámbito, pues, este fenómeno representa una expansión cualitativa del sistema de consecuencias jurídicas, que debe mantener coherencia con principios fundamentales como la proporcionalidad, humanidad y orientación resocializadora de las penas.

b. Otras manifestaciones

La incidencia de la inteligencia artificial en el derecho penal no se limita solo a las categorías fundamentales ya analizadas, sino que permea otras instituciones dogmáticas de significativa relevancia. En primer lugar, la teoría del error requiere reconsideraciones sustanciales en contextos algorítmicamente mediados; los tradicionales error de tipo y error de prohibición adquieren dimensiones particulares cuando la comprensión de elementos fácticos o normativos está mediada por sistemas de IA. La vencibilidad del error, tradicionalmente vinculada a la capacidad y deber de información del sujeto, debe renovarse considerando la opacidad algorítmica y la delegación cognitiva que caracteriza a muchos sistemas inteligentes. ¿Cómo valorar, por ejemplo, el error sobre los elementos normativos cuando él proviene de interpretaciones realizadas por algoritmos de procesamiento del lenguaje natural, cuyo funcionamiento interno resulta insondable incluso para sus programadores?

En segundo término, las categorías vinculadas al *iter criminis* experimentan una transformación significativa. La secuencia tradicional de ideación, preparación, ejecución y consumación se difumina en entornos donde la programación, el entrenamiento y la implementación de un sistema de IA pueden constituir fases separadas temporalmente y ejecutadas por distintos actores; la determinación del momento exacto en que se inicia la ejecución típica, crucial para la punibilidad de la tentativa, resulta particularmente problemática cuando el desarrollo del plan criminal implica despliegues graduales de sistemas semiautónomos.

La teoría de la autoría y participación, por su parte, requiere adaptaciones que trascienden la mera distribución de responsabilidades. Categorías como la coautoría, la autoría mediata y la complicidad deben reconsiderarse cuando el dominio del hecho se ejerce sobre o mediante sistemas con capacidad de aprendizaje y toma de

decisiones; resulta particularmente problemática la posible consideración de sistemas de IA como “instrumentos” en el sentido de la autoría mediata, especialmente cuando estos poseen grados significativos de autonomía decisional. Asimismo, la aplicación de criterios como el acuerdo común, la división de funciones o la aportación esencial en fase ejecutiva –propios de la coautoría– requiere adaptaciones en contextos donde parte de la “distribución de roles” se realiza con componentes artificiales.

El concurso de delitos presenta igualmente desafíos específicos. La unidad o pluralidad de acciones, criterio fundamental para distinguir entre concurso real e ideal, debe reconsiderarse cuando un mismo despliegue algorítmico puede generar simultáneamente múltiples resultados lesivos en diversos contextos espacio-temporales.

La jurisdicción y la aplicación territorial de la ley penal constituyen otro ámbito donde la IA genera tensiones significativas. La naturaleza intrínsecamente transnacional de muchos sistemas de IA, con componentes de desarrollo, entrenamiento, implantación y efectos distribuidos globalmente, desafía los principios tradicionales de territorialidad y personalidad; emergen así propuestas de modelos jurisdiccionales basados en criterios como el “control efectivo del sistema” o el “centro de intereses algorítmicos”, que buscan superar las limitaciones de los principios tradicionales.

En cuanto a la responsabilidad por omisión, el análisis de la posición de garante requiere adaptaciones significativas; la delegación de funciones de supervisión, control o intervención en sistemas de IA genera interrogantes sobre la configuración y límites del deber de vigilancia. La automatización de procesos decisionales, adviértase, no elimina sino que reconfigura la posición de garante de quienes mantienen poder fáctico y jurídico sobre sistemas potencialmente lesivos; esta reconsideración resulta particularmente relevante en ámbitos como la seguridad informática, la medicina algorítmica o los sistemas financieros automatizados.

Finalmente, la prueba digital y sus implicaciones para las garantías procesales, aunque pertenece formalmente al ámbito procesal, mantiene una estrecha relación con la configuración material del derecho penal; la capacidad probatoria de evidencias generadas por sistemas de IA, su fiabilidad epistémica y su compatibilidad con garantías como la presunción de inocencia inciden directamente en la operatividad práctica de categorías dogmáticas como el dolo, la imprudencia o el nexo causal (Nieva Fenoll, 2023).

c. Conclusión

La incidencia de la inteligencia artificial en las construcciones teóricas del derecho penal, parte general, no constituye un fenómeno periférico, sino una transformación que afecta al núcleo mismo de esta disciplina. Las categorías conceptuales tradicionales –acción, tipicidad, antijuridicidad, culpabilidad, bien jurídico, pena,

etc.– experimentan renovaciones profundas que requieren una adaptación constructiva de la dogmática penal.

Lejos de adoptar posiciones tecnofóbicas o tecnofílicas extremas, la ciencia penal contemporánea debe desarrollar un enfoque equilibrado que permita, por un lado, preservar garantías y principios fundamentales del derecho penal y, por otro, adaptar sus categorías conceptuales para responder adecuadamente a los desafíos que plantea la IA. Esta adaptación exige un diálogo interdisciplinario fructífero entre juristas, científicos computacionales, filósofos de la tecnología y éticos, que permita desarrollar construcciones dogmáticas coherentes con la realidad tecnológica actual y futura. Solo mediante este esfuerzo colaborativo será posible preservar la función del derecho penal como instrumento de control social legítimo en la era digital.

Cabe destacar que esta ampliación de perspectiva confirma que la transformación del derecho penal ante la IA no constituye un fenómeno aislado o sectorial, sino una reconfiguración sistémica que afecta transversalmente a la totalidad de las categorías e instituciones dogmáticas, requiriendo un esfuerzo intelectual que mantenga la coherencia interna del sistema mientras incorpora las particularidades de los entornos tecnológicamente mediados.

En fin, en síntesis, con Seher (2016), se puede concluir lo siguiente: en primer lugar, las inteligencias artificiales no pueden “actuar” en el sentido penal porque les falta la capacidad de percibir normas como órdenes por sí mismos y, por lo tanto, ser directamente interpeladas por ellas; el conocimiento de normas lo obtienen solo a través de la mediación de programadores humanos. Por lo tanto, no pueden ser destinatarios de normas jurídicas.

En segundo lugar, esas inteligencias no pueden ser declaradas culpables porque en el momento de su acción dañina no podían actuar de otra manera que como lo hicieron; declararlas culpables tampoco tiene sentido desde una perspectiva funcionalista, porque la comunidad jurídica no interpreta sus acciones como violaciones de normas. En tercer lugar, tampoco pueden ser castigadas porque mediante una imposición de desventajas hacia ellas no serviría a ningún fin reconocido de la pena.

Adicional a ello, en cuarto lugar, los operadores de esas inteligencias pueden hacerse penalmente responsables por acciones erróneas dentro de sus ámbitos de responsabilidad jurídica; el límite de tal punibilidad será probablemente establecido por el órgano legislativo mediante procedimientos de autorización que abran a los operadores un espacio libre de pena del riesgo permitido. Y, para culminar, en quinto lugar, quien no puede actuar en el sentido jurídico, quien no puede ser declarado culpable y quien no es un destinatario idóneo para una condena y un castigo, no es una “persona” en el derecho. Por tanto, las inteligencias artificiales seguirán siendo considerados como máquinas o herramientas, nunca sujetos de derecho penal.

IV. LA IA EN EL PROCESO PENAL

Para empezar se debe decir que la IA, desde hace rato, se utiliza en el proceso penal a veces de manera rudimentaria y en otras de forma avanzada (Nieva, 2022, p. 19). Sin embargo, la evaluación de riesgos procesales –principalmente el peligro de fuga y la potencial obstaculización de la investigación– constituye un elemento central en decisiones críticas como la imposición de prisión preventiva, la determinación de fianzas o la aplicación de medidas cautelares alternativas. Tradicionalmente basada en apreciaciones jurídicas sustentadas en la experiencia y el criterio judicial, esta evaluación ha experimentado una profunda transformación con la introducción de sistemas algorítmicos que prometen mayor objetividad, consistencia y precisión predictiva. Esta transición hacia una “justicia actuarial” representa uno de los campos más controvertidos y consecuentes de la aplicación de IA en el proceso penal, con implicaciones directas para los derechos fundamentales como la libertad personal, la presunción de inocencia y la igualdad ante la ley.

Los sistemas de evaluación algorítmica de riesgos procesales ahora utilizados presentan diversas arquitecturas y enfoques metodológicos; así los diseños basados en puntajes estáticos incluyen el PSA (*Public Safety Assessment* o *Evaluación de Seguridad Pública*), desarrollado por la Laura and John Arnold Foundation e introducido en numerosas jurisdicciones estadounidenses. Este sistema analiza nueve factores objetivos (edad, historial criminal previo, comparecencias anteriores, etc.) para generar tres puntuaciones distintas: riesgo de no comparecencia, peligro de reincidencia general y riesgo de reincidencia violenta. Otro ejemplo, es el VPRAI (Virginia Pretrial Risk Assessment Instrument o Instrumento de Evaluación de Riesgo Previo al Juicio de Virginia), que evalúa ocho factores de riesgo para clasificar a los acusados en seis niveles distintos de riesgo procesal, con recomendaciones específicas de supervisión para cada nivel.

Los modelos basados en *machine learning* incluyen el COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*: Perfilado para la Gestión de Delinquentes en Correccionales para Sanciones Alternativas), desarrollado por Northpointe (ahora Equivant), que utiliza algoritmos propietarios de *machine learning* para evaluar riesgos, en atención a factores criminológicos tanto estáticos como dinámicos, incluidas variables socioeconómicas y de entorno.

Por ejemplo, el HART (*Harm Assessment Risk Tool*), empleado por la policía de Durham (Reino Unido), utiliza *random forests* (bosques aleatorios) para clasificar acusados en categorías de riesgo basándose en 34 variables, con inclusión de su historial criminal, las características demográficas y los códigos postales, etc. A su vez, los sistemas de última generación incluyen modelos de *deep learning*, que utilizan redes neuronales para identificar patrones complejos no lineales en datos históricos. El sistema experimental DEEPJUDGE, desarrollado por investigadores del Instituto Tecnológico de Massachussets (MIT), incorpora análisis de lenguaje na-

tural de documentos judiciales para extraer factores predictivos previamente no considerados.

También existen sistemas de Inteligencia Artificial Explicable (XAI) como CORELS (Certifiably Optimal Rule Lists), desarrollado por investigadores de Harvard; este sistema genera reglas de decisión transparentes y matemáticamente verificables, lo que permite entender claramente cómo llega a sus conclusiones. CORELS hace un compromiso consciente: sacrifica cierto nivel de precisión predictiva para ganar en interpretabilidad, permitiendo que jueces, abogados y otros profesionales del derecho puedan comprender y evaluar los razonamientos utilizados. Esta característica es especialmente valiosa en el ámbito jurídico-penal, donde la transparencia en la toma de decisiones resulta fundamental para garantizar el debido proceso y el empleo adecuado y razonable de los datos personales

La introducción de estos sistemas ha generado un creciente *corpus* de investigación evaluativa con resultados mixtos. Entre los estudios favorables, una evaluación del PSA en Kentucky (Pretrial Safety Assessment o Evaluación de Seguridad Previa al Juicio) [una herramienta de evaluación de riesgos desarrollada por la Fundación Laura y John Arnold, consistente en un algoritmo que ayuda a los jueces y a otros funcionarios del sistema judicial a tomar decisiones más informadas sobre si un acusado debe ser liberado o detenido antes de su juicio] verificó reducciones del 15% en tasas de detención preventiva sin incrementos significativos en no comparecencias o nuevos arrestos (Stevenson, 2018). La introducción del modelo VPRAI en Virginia se asoció con una reducción del 70% en la detención preventiva de acusados identificados como de “bajo riesgo”, manteniendo tasas de comparecencia superiores al 94% (VanNostrand & Rose, 2009). a su vez, un meta-análisis de Desmarais y Singh (2013) concluyó que los instrumentos estructurados de evaluación de riesgos superan consistentemente las predicciones puramente clínicas o intuitivas en términos de precisión predictiva.

Sin embargo, estudios críticos como la investigación de PROPUBLICA –una organización de periodismo de investigación sin fines de lucro fundada en 2007 en Estados Unidos que se especializa en reportajes de interés público con un enfoque en abusos de poder, corrupción y temas donde el interés público está en riesgo– (Angwin *et al.*, 2016) sobre COMPAS, muestran tasas significativamente distintas de falsos positivos entre acusados afroamericanos (44.9%) y caucásicos (23.5%), sugiriendo sesgos sistemáticos.

En este contexto, los trabajos de Dressel y Farid (2018) enseñan que personas sin experiencia jurídica pueden igualar la precisión predictiva de COMPAS (en un 65%) simplemente evaluando siete factores básicos, lo que cuestiona así el valor añadido de sistemas complejos. Incluso, un estudio longitudinal en California (Stevenson & Doleac, 2021) estableció que los jueces tendían progresivamente a alinearse con recomendaciones algorítmicas, incluso cuando estas contradecían su criterio inicial, fenómeno denominado “deferencia algorítmica”. Esta evidencia mixta refleja

las complejas interacciones entre el diseño algorítmico, la introducción práctica, los factores institucionales y el comportamiento judicial; se debe, pues, hacer frente a un riesgo cierto, “que no es otro que el peligro de que el juez tienda a acomodar sus decisiones, sin apenas más valoraciones adicionales, a lo que resulte de aplicar el sistema IA pertinente al asunto que debe resolver” (Hoyos Sancho, 2020, p. 36).

La utilización de estos sistemas genera, pues, profundas tensiones con principios fundamentales del proceso penal; la evaluación de riesgos fundada en características personales o socioeconómicas puede constituir una forma de “pre-castigo” basado en correlaciones estadísticas, no en hechos probados. Como señaló el Tribunal Constitucional alemán en su decisión sobre perfilado preventivo (1 BvR 518/02), “los ciudadanos no deben ser tratados como objetos de vigilancia sin que hayan dado lugar a ello”; y, como ha dicho la doctrina, la vigilancia total construida “mediante algoritmos que dirigen tecnologías que, a su vez, impiden directamente la infracción”, supone “una restricción de la libertad mediante coacción física, sin infracción previa y sin proceso debido”, que “constituye en realidad una «medida de seguridad predelictiva» con capacidad de generalización y tendencia expansiva” (Silva Sánchez, 2025, p. 140).

A ello se debe añadir que la estandarización algorítmica puede entrar en conflicto con el principio de que cada caso debe evaluarse según sus circunstancias particulares; por ejemplo, el Tribunal Supremo español ha reiterado (STS 308/2018) que las decisiones sobre medidas cautelares deben “atender a las circunstancias específicas de cada caso concreto”, principio potencialmente amenazado por sistemas que generalizan basándose en correlaciones estadísticas (sobre ello, Planchadell Gargallo, 2024).

Además, la complejidad algorítmica plantea desafíos para la obligación judicial de fundamentar decisiones restrictivas de derechos. El caso emblemático “State v. Loomis” (2016; Martínez Garay, 2018) abordó esta tensión, estableciendo que los jueces pueden utilizar evaluaciones algorítmicas como factor complementario, pero no como fundamento único o principal, debiendo explicitar las consideraciones independientes que sustentan su decisión. Adicional a ello, la opacidad técnica y la frecuente protección como secreto comercial de estos algoritmos obstaculiza el ejercicio efectivo del derecho de defensa; el Caso “Houston Federation of Teachers v. Houston Independent School District” (2017) estableció un precedente importante en esta materia al determinar que sistemas de evaluación algorítmica cuya metodología permanezca inaccesible podrían violar derechos procesales fundamentales.

La incorporación de variables correlacionadas con características protegidas (raza, género, estatus socioeconómico) puede perpetuar y legitimar técnicamente discriminaciones históricas. Nuestra Corte Constitucional (Sentencia T-043/20) advirtió, específicamente, sobre el riesgo de “blanqueo de discriminación” mediante algoritmos aparentemente neutrales; incluso, ese mismo organismo mediante Sentencia T-323 del cuatro de agosto de 2024, emitió una serie de lineamientos en ma-

teria del uso de la IA por parte de la judicatura e invitó a realizar prácticas adecuadas en esta materia para velar por la seguridad jurídica y un uso ético y responsable de estas herramientas.

Ante estos desafíos, diversas jurisdicciones han comenzado a desarrollar marcos normativos específicos. La de California (SB 10, aunque posteriormente derogada), por ejemplo, señaló que las evaluaciones algorítmicas pueden utilizarse únicamente como herramientas informativas complementarias, prohibiendo decisiones automáticas basadas exclusivamente en puntuaciones de riesgo. La de Nueva York introdujo la “AI-LEAD Act” (*Artificial Intelligence-Limiting Experimental Algorithmic Decisions*), que prohíbe el uso de sistemas de evaluación protegidos como secreto comercial en decisiones sobre libertad condicional, exigiendo divulgación completa de variables, ponderaciones y metodologías.

La de Pennsylvania, a su turno, estableció un comité de supervisión independiente que evalúa anualmente el impacto de su *Risk Assessment Tool* (Herramienta de Evaluación de Riesgos) incluyendo análisis específicos de potenciales disparidades raciales o socioeconómicas. Menciónese también que la justicia de Nueva Jersey prohibió el uso de códigos postales, vecindario y otras variables geográficas en su sistema PSA –recuérdese: *Public Safety Assessment* o *Evaluación de Seguridad Pública*–, tras evidenciar que funcionaban como proxies (indicadores) de composición racial. Incluso, la Unión Europea, mediante el artículo 22.1. del RGPD (Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo de 27 de abril de 2016), señala el derecho a no estar sujeto a decisiones basadas únicamente en procesamiento automatizado cuando se producen efectos jurídicos significativos, garantizando intervención humana sustantiva: “Todo interesado tendrá derecho a no ser objeto de una decisión basada únicamente en el tratamiento automatizado, incluida la elaboración de perfiles, que produzca efectos jurídicos en él o le afecte significativamente de modo similar”.

La experiencia acumulada sugiere directrices para un desarrollo jurídicamente adecuado en el ámbito penal de tal forma que los sistemas algorítmicos se conciban como fuentes informativas adicionales y no como sustitutos del criterio judicial fundamentado. Todo sistema introducido en contextos procesales penales debe ser completamente auditable, incluyendo acceso a variables, ponderaciones y metodologías específicas. Es necesario, pues, el uso de mecanismos de monitoreo permanente que evalúen resultados reales, detectando disparidades que puedan ser discriminatorias.

Es fundamental que los jueces, fiscales y otros funcionarios judiciales reciban capacitación obligatoria sobre cómo interpretar las predicciones algorítmicas; esto incluye entender las limitaciones estadísticas de estos sistemas y cómo pueden estar sesgados. Por ejemplo, los algoritmos que predicen la reincidencia a menudo tienen tasas de error más altas para ciertos grupos raciales, lo que puede llevar a decisiones injustas. Es importante, entonces, el debido desarrollo de derechos procesales específicos mediante la evolución normativa para establecer explícitamente garantías

adaptadas al contexto tecnológico, con la inclusión del sagrado derecho a contradicción algorítmica efectiva.

La evaluación de riesgos basada en algoritmos es un ejemplo claro de la tensión que genera la IA en el derecho penal. Por un lado, puede mejorar la toma de decisiones al procesar grandes cantidades de información; por otro, plantea desafíos para principios fundamentales como la presunción de inocencia, la igualdad ante la ley y la necesidad de que cada caso sea tratado de manera individual. Su adecuada regulación constituye, pues, uno de los desafíos más urgentes para preservar la legitimidad del proceso penal en la era algorítmica.

Los Sistemas de Apoyo a la Decisión Judicial (SADJ) basados en la inteligencia artificial representan una de las aplicaciones más ambiciosas y potencialmente transformadoras de la tecnología en el ámbito procesal penal. Estos sistemas utilizan algoritmos avanzados para asistir a jueces y magistrados en diversas dimensiones del proceso decisorio, desde la aplicación consistente de criterios legales hasta la estimación de consecuencias probables de diferentes alternativas sentenciadoras. De ahí que su empleo plantee una profunda reconsideración de la función jurisdiccional tradicional, y ponga en tensión conceptos fundamentales como la independencia judicial (Gómez Colomer, 2023), la individualización de la respuesta penal y la responsabilidad por decisiones que afectan directamente a los derechos fundamentales.

Ahora bien, téngase en cuenta que esos SADJ contemporáneos abarcan un espectro amplio de capacidades y diseños arquitectónicos. Los sistemas de recomendación sentenciadora incluyen algoritmos predictivos de resultados como el ya citado COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*) en su función sentenciadora, que analiza variables criminológicas para predecir probabilidades de reincidencia y sugerir intervenciones calibradas a perfiles de riesgo específicos. Además, los modelos de consistencia judicial, como el JUDGE (*Judicial Uniformity through Data-Guided Evaluations*) implantado experimentalmente en tribunales federales estadounidenses, analizan patrones históricos de sentencias para casos similares, con la identificación de casos atípicos (*outliers*) y la promoción de la uniformidad jurisdiccional. Además, las calculadoras de proporcionalidad, como el sistema NEO (*Netherlands Sentencing Calculator*), cuantifican la gravedad delictiva mediante ponderación algorítmica de factores legalmente relevantes, sugiriendo rangos proporcionales de penalidad.

También, los sistemas de análisis jurisprudencial avanzado incluyen motores de búsqueda jurídica potenciados por IA, como el ROSS Intelligence y el Lex Machina, que utilizan procesamiento de lenguaje natural para identificar precedentes relevantes, patrones decisorios y argumentos judiciales aplicables a casos concretos. Los sistemas predictivos jurisprudenciales, como PREDICTICE (Francia) y LEGAL ANALYTICS (Thomson Reuters), analizan patrones históricos en jurisdicciones específicas para estimar probabilidades de diferentes resultados según características del caso. Además, los extractores automáticos de *ratio decidendi*, como JURIMPACT, identifican

automáticamente principios jurídicos subyacentes en decisiones previas, facilitando su aplicación consistente.

De igual manera, los asistentes de razonamiento judicial incluyen verificadores de consistencia lógica, como LOGICCHECK, que analizan argumentaciones judiciales para identificar potenciales contradicciones internas, falacias lógicas o saltos argumentativos. También, los evaluadores de suficiencia probatoria, como EVIDENCER, aplican estándares bayesianos –recuérdese el teorema de Bayes, fundamento de estos estándares, que permite actualizar la probabilidad de una hipótesis a medida que se incorpora nueva evidencia– para evaluar la solidez estadística de conclusiones derivadas de material probatorio disponible. Y, añádase, los constructores de argumentación, como ARGUBUILDER, generan estructuras argumentativas alternativas basadas en diferentes teorías jurídicas aplicables.

La adopción de un SADJ presenta variaciones significativas entre jurisdicciones. Estados Unidos, recuérdese, lidera el uso con sistemas como COMPAS, ampliamente utilizado en jurisdicciones estatales para decisiones sentenciadoras y de libertad condicional. El sistema E-CODIS (*Electronic Court Disposition Input System*) en California ha procesado más de 350,000 casos, generando recomendaciones en aproximadamente el 62% de sentencias menores. Estudios longitudinales (Stevenson, 2018) sugieren que las tasas de coincidencia entre recomendaciones algorítmicas y decisiones judiciales efectivas aumentan progresivamente con el tiempo, indicando creciente influencia algorítmica.

Asimismo, Estonia ha desarrollado el sistema más avanzado de Europa con su *Robot Judge*, actualmente operativo para casos de cuantía menor, que analiza documentación y genera propuestas de resolución. Para el 2024 pasado, se planeaba expandir su ámbito a delitos menores con consentimiento explícito de las partes. China, a su vez, ha implementado el sistema “206” en más de 3.500 tribunales, que analiza más de 100 variables caso por caso, incluyendo datos biométricos del acusado durante audiencias (microexpresiones faciales, respuestas fisiológicas) para generar recomendaciones vinculadas a precedentes. Las estadísticas oficiales reportan un 85% de concordancia entre recomendaciones y sentencias efectivas.

Brasil, por su parte, ha implantado el sistema VICTOR por parte del Supremo Tribunal Federal, que procesa automáticamente recursos de admisibilidad, identificando precedentes aplicables y verificando requisitos procesales, con lo cual ha reducido el tiempo medio de procesamiento inicial de 44 a 5 días. La Unión Europea, a su vez, presenta un enfoque más cauteloso, con proyectos experimentales como PROMETEA (Tribunal Administrativo de París) y JURIMETRÍA (España), que funcionan principalmente como asistentes de investigación jurídica sin capacidad recomendatoria directa.

La evidencia empírica sobre la efectividad de estos instrumentos muestra resultados contrastantes. Kleinberg *et al.* (2018), por ejemplo, mostraron que algoritmos

relativamente simples superaban consistentemente a jueces humanos en predicción de riesgo de incomparecencia, con reducciones potenciales de entre 30-42% en tasas de detención preventiva manteniendo riesgos constantes. Simultáneamente, el análisis retrospectivo de Dressel & Farid (2018) sobre COMPAS cuestionó el valor añadido de sistemas complejos, demostrando que personas sin formación jurídica con la utilización de solo siete variables básicas igualaban su precisión predictiva (65%). El estudio longitudinal de Green & Chen (2021) identificó un fenómeno preocupante de “conformismo algorítmico”: los jueces tendían progresivamente a alinearse con recomendaciones automatizadas incluso cuando éstas contradecían su criterio inicial, especialmente en casos ambiguos o complejos.

La implementación de un SADJ genera tensiones constitucionales y procesales profundas. La independencia judicial, concepto constitucionalmente protegido en prácticamente todas las democracias modernas, potencialmente se ve amenazada por sistemas que pueden ejercer una sutil presión conformista; como lo declaró el Tribunal Constitucional español (STC 12/2020), la independencia judicial “no admite más sumisión que al imperio de la ley”, principio complejo de reconciliar con una autoridad algorítmica percibida como “científica”.

Además; no se olvide, la motivación judicial efectiva es un requisito procesal fundamental potencialmente comprometido cuando las decisiones se fundamentan en *outputs* algorítmicos no completamente comprensibles por el propio juzgador; al respecto, el Tribunal Europeo de Derechos Humanos ha establecido reiteradamente (STEDH Taxquet c. Bélgica, 2010) que toda decisión penal debe explicitar “de manera clara e indudable los fundamentos en que se basa”, estándar difícil de satisfacer mediante delegación parcial de la tarea juzgadora en sistemas opacos.

La individualización judicial, principio central del derecho penal moderno que exige consideración específica de las circunstancias particulares, potencialmente se ve erosionada por sistemas diseñados para identificar patrones generales y promover consistencia estadística. Así, la humanidad de la justicia, dimensión cualitativa fundamentada en nociones como equidad, compasión y contextualización social, constituye aspectos difícilmente cuantificables algorítmicamente. El Tribunal Constitucional italiano (Sentenza 253/2003) enfatiza que la justicia penal “no puede reducirse a (una) operación mecánica de aplicación normativa” debiendo incorporar “valoraciones humanas inherentemente discrecionales”.

La transparencia en el uso de algoritmos es esencial para garantizar el derecho de defensa. Sin embargo, muchos sistemas de IA están protegidos como secretos comerciales, lo que dificulta que los abogados y los jueces entiendan cómo se toman las decisiones. Para solucionar esto, se podrían implementar leyes que exijan a las empresas compartir los códigos fuente de los algoritmos utilizados en procesos judiciales, siempre que no comprometan la propiedad intelectual. El emblemático caso “Wisconsin v. Loomis” (2016) ilustró esta tensión, cuando la defensa no pudo acceder al funcionamiento interno del COMPAS por constituir secreto comercial.

Ante estos desafíos, diversas jurisdicciones y sistemas penales han comenzado a desarrollar marcos normativos específicos. Instrumentos como la *Carta Ética Europea sobre el Uso de la Inteligencia Artificial en los Sistemas Judiciales* (CEPEJ, 2018) establecen que los SADJ deben limitarse a funciones informativas y organizativas, prohibiendo delegación decisoria efectiva. Francia, por ejemplo, modificó su Código de Procedimiento Penal (artículo L10-1) para prohibir explícitamente el uso de SADJ cuya metodología no sea completamente transparente y auditable, incluyendo el acceso defensivo pleno. La *Ley de Algoritmos Justos* de Washington exige que todo SADJ implantado en contextos judiciales sea diseñado incorporando mecanismos de explicabilidad, documentación completa de decisiones de diseño y evaluaciones pre-implementación de impacto discriminatorio.

A su turno, el *Consejo General del Poder Judicial* español estableció en 2020 un *Comité de Ética Judicial Digital* con capacidad vinculante para evaluar y autorizar cualquier implementación de SADJ en tribunales nacionales. Además, varios estados estadounidenses han implantado modelos de “diseño participativo” en cuya virtud representantes de la fiscalía, la defensa, la judicatura y la sociedad civil intervienen obligatoriamente en el desarrollo, la parametrización y la evaluación continua de SADJ. Para no ir muy lejos, el Consejo Superior de la Judicatura colombiano adoptó el 16 de diciembre de 2024 el Acuerdo PCSJA24-12243, mediante en cual se trazan diversas directrices para el uso de la IA en la Rama Judicial nacional.

El análisis comparado sugiere, entonces, el respeto a unos principios fundamentales para un uso compatible con garantías básicas. Es necesaria, pues, la subsidiariedad algorítmica, con la conceptualización explícita del SADJ como herramientas estrictamente informativa, nunca sustitutiva del criterio judicial fundamentado. Su función debe ser expandir información disponible, no reemplazar la valoración humana. Se requiere, pues, transparencia metodológica completa, con el rechazo absoluto de algoritmos “caja negra” protegidos como propiedad intelectual. Todo sistema usado debe ser completamente auditable, incluyendo acceso a los códigos fuente, variables consideradas, ponderaciones específicas y conjuntos de entrenamiento.

Además, la contradicción algorítmica efectiva implica el desarrollo de mecanismos procesales específicos que permitan a la defensa cuestionar efectivamente las recomendaciones algorítmicas, incluyendo acceso a expertos técnicos financiados públicamente. Asimismo, es importante la prohibición de *inputs* problemáticos, con exclusión explícita de variables que funcionen como proxies de características constitucionalmente protegidas como raza, género o estatus socioeconómico. La evaluación de impacto continua requiere el empleo de sistemas de monitoreo permanente que evalúen patrones decisorios resultantes, identificando potenciales efectos discriminatorios, conformismos algorítmicos o desviaciones de principios penológicos fundamentales.

La formación judicial especializada implica capacitación obligatoria y continua para servidores judiciales sobre el funcionamiento, las limitaciones y los potenciales

sesgos de los sistemas adoptados, promoviendo el uso crítico e informado. La responsabilidad judicial explícita requiere mecanismos que refuercen la comprensión de que la responsabilidad decisoria recae exclusivamente en el juzgador humano, incluyendo obligaciones de explicitar su razonamiento independiente cuando se aparta o confirma recomendaciones algorítmicas.

En fin, los sistemas de apoyo a la decisión judicial representan simultáneamente una oportunidad para mejorar la consistencia, la eficiencia y la fundamentación empírica del proceso decisorio, y un desafío fundamental para conceptos nucleares de la función jurisdiccional en democracias constitucionales. Su adecuada regulación, que preserve beneficios potenciales mientras salvaguarde principios fundamentales, constituye uno de los principales retos jurídicos del siglo XXI, con implicaciones directas para la legitimidad y aceptabilidad social del sistema penal.

V. LA IA EN LA INVESTIGACIÓN CRIMINAL

En esta precisa materia son diversos los escenarios en los cuales cobra mucha importancia esta herramienta; obsérvense los cuatro más importantes:

a. Herramientas de reconocimiento facial

Este instituto representa una de las aplicaciones más extendidas y controvertidas de la inteligencia artificial en el ámbito de la investigación criminal; esta tecnología utiliza algoritmos de aprendizaje profundo, principalmente redes neuronales convolucionales (CNN), para identificar o verificar la identidad de una persona analizando y comparando patrones faciales. Su empleo en el sistema de justicia penal ha transformado radicalmente las capacidades investigativas de los cuerpos policiales, pero también ha generado profundas preocupaciones sobre derechos fundamentales.

Los sistemas de reconocimiento facial operan mediante un proceso secuencial que incluye la detección facial, en cuya virtud el algoritmo localiza rostros en una imagen o video; la alineación facial, que normaliza la imagen para compensar variaciones de posición, iluminación o expresión; la extracción de características, que identifica y cuantifica rasgos distintivos, generando una "huella facial" digital única; y, finalmente, la comparación de patrones, que confronta esta huella con una base de datos para identificar coincidencias potenciales.

Su uso en contextos policiales abarca la identificación en tiempo real mediante cámaras de vigilancia conectadas a bases de datos para identificar sospechosos o personas buscadas en espacios públicos; el análisis forense retrospectivo aplicado a grabaciones o fotografías relacionadas con hechos delictivos; el control de acceso a instalaciones sensibles, limitando el ingreso a personas autorizadas; y la verifi-

cación de identidad, confirmando la identidad de detenidos o comparecientes en procesos penales.

Junto al reconocimiento facial, otros sistemas biométricos relevantes incluyen el reconocimiento del iris y la retina, que ofrece mayor precisión pero requiere cooperación del sujeto; la identificación por voz, particularmente útil para vincular comunicaciones anónimas con sospechosos conocidos; el análisis de marcha, que identifica patrones únicos en la forma de caminar; y, añádase, el reconocimiento de venas, que analiza el patrón vascular subcutáneo.

Como es obvio, el despliegue de estas tecnologías varía significativamente entre sistemas jurídicos y jurisdicciones. China, por ejemplo, ha introducido la red más extensa de reconocimiento facial del mundo, con más de 200 millones de cámaras conectadas a sistemas de IA, utilizándola extensivamente para la identificación y el seguimiento de sospechosos, el refuerzo del orden público y la prevención delictiva, incluso ahora desde satélites situados en el espacio; el sistema denominado “Skynet” puede localizar a individuos entre la población en cuestión de minutos. A su turno, en Estados Unidos se presenta un panorama fragmentado en atención a que agencias federales como el FBI (con su sistema *Next Generation Identification*) y estatales utilizan ampliamente estas tecnologías, mientras que ciudades como San Francisco, Boston y Portland han prohibido su uso por preocupaciones relacionadas con los derechos civiles.

Asimismo, en el Reino Unido se han desplegado sistemas de reconocimiento facial en tiempo real en espacios públicos de Londres y otras ciudades principales, generando controversia jurídica tras decisiones judiciales que cuestionan su proporcionalidad. A su turno, la Unión Europea, con su Reglamento General de Protección de Datos (RGPD), considera los datos biométricos como categorías especiales de datos personales, e impone restricciones significativas para su procesamiento. De forma complementaria, la Ley de Inteligencia Artificial (AI Act) de la Unión Europea aprobada oficialmente en marzo de 2024 clasifica ciertos usos de reconocimiento biométrico como de “alto riesgo” e impone requisitos estrictos, e incluso prohíbe su uso en tiempo real en espacios públicos salvo excepciones limitadas. En América Latina, países como Brasil, Argentina y México han comenzado a emplear estas tecnologías en aeropuertos, estaciones de transporte y zonas de alta criminalidad, generalmente sin marcos regulatorios específicos para estas aplicaciones, rigiéndose principalmente por leyes generales de protección de datos cuando estas existen, lo que obviamente genera dificultades cuando se piensa en modelos de persecución penal ajustados a los dictados propios de un Estado de derecho liberal y democrático.

Por eso, como es de esperarse, la introducción de sistemas de reconocimiento facial en la investigación criminal plantea múltiples desafíos jurídicos. En primer lugar, surge la preocupación por el derecho a la privacidad, dado que la vigilancia masiva indiscriminada mediante estas tecnologías constituye una injerencia significativa en la privacidad; a este respecto, el Tribunal Europeo de Derechos Humanos, en el caso

“S. y Marper c. Reino Unido” (2008), estableció que la mera recopilación y almacenamiento de datos biométricos sin consentimiento representa una interferencia en la vida privada protegida por el artículo 8° del Convenio Europeo.

Otra preocupación importante es la presunción de inocencia, porque la vigilancia algorítmica previa a cualquier indicio de criminalidad invierte la lógica tradicional de la investigación penal, sometiendo a escrutinio continuo a ciudadanos sin sospecha individualizada. También existe el riesgo de discriminación, pues múltiples estudios científicos (Buolamwini & Gebru, 2018; NIST, 2019) demuestran tasas de error significativamente mayores en la identificación de personas de piel oscura, mujeres y jóvenes, planteando inaceptables riesgos de discriminación algorítmica. A no dudarlo, los sistemas de reconocimiento facial tienen tasas de error más altas para ciertos grupos demográficos, lo que puede resultar en falsos positivos (identificar incorrectamente a una persona) o falsos negativos (no identificar a una persona que debería ser reconocida).

Asimismo, el debido proceso y el derecho de defensa pueden verse comprometidos debido a que la complejidad técnica y frecuente opacidad de estos sistemas dificulta su cuestionamiento efectivo por parte de la defensa, potencialmente vulnerando el principio de igualdad de armas. Asimismo, la conciencia de estar bajo vigilancia constante puede producir un “efecto intimidatorio” (*chilling effect*) sobre el ejercicio de derechos como la libertad de expresión, reunión y manifestación. Por último, surgen cuestiones sobre la integridad, autenticidad y fiabilidad de la evidencia obtenida mediante estos sistemas, incluyendo problemas de alteración, falsificación (*deepfakes*) y valoración judicial.

Así las cosas, para abordar estos desafíos, diversos sistemas penales, jurisdicciones y organismos internacionales han propuesto marcos regulatorios que incluyen la limitación en atención a la gravedad delictiva y restringen su uso a la investigación de delitos graves tipificados de forma taxativa. Además, se dispone el control judicial previo mediante la autorización judicial fundamentada para su introducción, de forma similar a la intervención de las comunicaciones; también se prevé la transparencia algorítmica y se exige que los sistemas utilizados en contextos penales sean auditables por expertos independientes y comprensibles para los afectados.

Igualmente, obsérvese, se dispone la supervisión independiente a través de organismos especializados para autorizar, monitorear y evaluar el uso de estas tecnologías; se introducen garantías técnicas mediante la imposición de estándares mínimos de precisión, validados periódicamente a través de pruebas rigurosas; se introducen limitaciones temporales y territoriales, y se limita su despliegue a periodos y zonas específicas, con lo cual se evita la normalización de la vigilancia masiva permanente; y, en fin, se introducen las evaluaciones de impacto obligatorias y se exigen análisis previos sobre implicaciones para los derechos fundamentales y los riesgos discriminatorios que se corren.

El reconocimiento facial y otras tecnologías biométricas ilustran, pues, el dilema central que plantea la IA en el ámbito penal, esto es, equilibrar su indudable potencial para mejorar la eficacia investigativa con la necesaria protección de derechos fundamentales y los principios jurídicos que definen un sistema penal legítimo, en una sociedad democrática. La regulación adecuada de estas tecnologías constituye, entonces, uno de los desafíos más urgentes para legisladores, tribunales y juristas en la actualidad.

b. El análisis predictivo

Este instrumento representa una de las aplicaciones más prometedoras y simultáneamente más controvertidas de la IA en el ámbito de la investigación criminal. Esta metodología, también conocida como “policía predictiva” (*predictive policing*), utiliza algoritmos avanzados para analizar grandes volúmenes de datos históricos con el objetivo de identificar patrones, predecir posibles focos delictivos y optimizar la asignación de recursos policiales. Desde luego, su desarrollo marca una transición significativa desde un enfoque principalmente reactivo hacia un paradigma proactivo y preventivo en la actuación policial.

El análisis predictivo criminal se fundamenta en la premisa de que la actividad delictiva, lejos de ser completamente aleatoria, sigue patrones discernibles influidos por variables espacio-temporales, socioeconómicas y situacionales. Los diseños actuales pueden clasificarse en tres categorías principales: en primer lugar, los *sistemas de predicción geoespacial* que están focalizados en predecir “dónde” y “cuándo” es más probable que ocurran determinados tipos de delitos. Estos sistemas, como lo son PREDPOL (ahora GEOLITICA) y HUNCHLAB, aplican algoritmos derivados de la sismología para modelar la propagación y concentración de la actividad criminal, bajo el supuesto de que ciertos delitos exhiben patrones de “contagio” similares a las réplicas sísmicas. Incluso, utilizan datos históricos sobre la ubicación precisa de incidentes previos, horas y días de la semana, características urbanísticas y arquitectónicas, patrones de movilidad poblacional, y factores ambientales como clima, iluminación o eventos especiales.

Los *sistemas de predicción centrados en personas*, en segundo lugar, están orientados a identificar individuos con mayor probabilidad de victimización o comisión de delitos; ejemplo de ello son el *Strategic Subject List* de Chicago (coloquialmente conocido como “*heat list*”: lista caliente) y el sistema HART (*Harm Assessment Risk Tool*) en Reino Unido. Estos modelos analizan antecedentes penales, relaciones sociales y afiliaciones, patrones de comunicación, historial de victimización, y variables socioeconómicas y demográficas.

Por su parte, en tercer lugar, los *sistemas de identificación de vínculos criminales* están diseñados para descubrir conexiones no evidentes entre incidentes aparentemente aislados, para lo cual utilizan técnicas de análisis de redes y minería de datos,

de cara a identificar patrones que sugieran la actividad de un mismo perpetrador o grupo criminal en múltiples casos.

Ahora bien, la utilización de sistemas predictivos ha generado resultados mixtos según diversos estudios empíricos. Los defensores de estos modelos citan investigaciones como la de Mohler *et al.* (2015), que documentó reducciones de entre 7.4% y 13.1% en delitos contra la propiedad en distritos de Los Ángeles donde se introdujo el PREDPOL, en comparación con zonas donde se mantuvieron métodos tradicionales de patrullaje. Similar reducción (el 23%, esta vez) fue reportada por el Departamento de Policía de Santa Cruz tras introducir el análisis predictivo geoespacial.

Sin embargo, estudios independientes como el de Lum & Isaac (2016) identificaron significativos problemas metodológicos, particularmente el “sesgo de retroalimentación”; en efecto, cuando la policía patrulla áreas identificadas por el algoritmo como de alto riesgo, naturalmente detecta más delitos en esas zonas y genera nuevos datos que refuerzan la predicción original y crean un ciclo de profecía autocumplida. Este fenómeno, adviértase, puede exacerbar la sobrevigilancia de comunidades ya marginalizadas.

Incluso, un meta-análisis conducido por el *Royal United Services Institute* (2020) concluyó que, si bien existe evidencia moderada de mejoras en la eficiencia táctica a corto plazo con la utilización de esta herramienta, la evidencia sobre impactos estratégicos a largo plazo en tasas de criminalidad general permanece inconsistente y metodológicamente problemática. Desde luego, el empleo del análisis predictivo en la investigación criminal plantea desafíos jurídicos fundamentales: existe una gran tensión con el principio de culpabilidad –y, por ende, con el principio del acto–, dado que los sistemas centrados en personas, al asignar perfiles de riesgo basados parcialmente en variables que el sujeto no controla (como el lugar de residencia o las conexiones sociales), parecen contradecir el principio fundamental de que la responsabilidad penal debe basarse exclusivamente en acciones humanas voluntarias.

También, surge preocupación respecto a la presunción de inocencia y a la estigmatización preventiva, pues la clasificación algorítmica de individuos como “de alto riesgo” previa a la realización de cualquier conducta delictiva, puede constituir una forma de estigmatización incompatible con la presunción de inocencia, principio rector del derecho penal moderno codificado en instrumentos como el artículo 14.2 del Pacto Internacional de Derechos Civiles y Políticos, incorporado a nuestro ordenamiento mediante la Ley 74 de 1968, y, por supuesto, en el artículo 29 de la Constitución Política y en múltiples desarrollos legales como los contenidos en la Ley 906 de 2004 o Código de Procedimiento Penal.

De igual manera, la discriminación algorítmica es otro problema significativo, pues múltiples investigaciones enseñan que estos sistemas frecuentemente replican y amplifican sesgos preexistentes en los datos históricos. Por ejemplo, estudio de PRO-PUBLICA (Angwin *et al.*, 2016; Staffler & Jany, 2020), demostró que los algoritmos de

predicción de riesgos clasificaban erróneamente a los acusados afroamericanos como de “alto riesgo” a casi el doble de la tasa que incluye a acusados caucásicos. Esta discriminación algorítmica potencialmente vulnera principios constitucionales como los ya señalados y el de igualdad ante la ley.

Las cuestiones propias de transparencia y “*due process*” (proceso debido) también son relevantes, porque la complejidad y la frecuente opacidad comercial de estos algoritmos (protegidos como secretos industriales) obstaculiza el escrutinio judicial y el efectivo ejercicio del derecho de defensa. El caso *State v. Loomis* (2016) constituye un precedente fundamental sobre el uso de la IA en el sistema judicial penal (Planchadell Gargallo, 2024); la Corte Suprema de Wisconsin avaló condicionalmente el uso del algoritmo predictivo COMPAS para evaluar el riesgo de reincidencia y dijo que no viola el debido proceso siempre que no sea el único factor determinante en la sentencia.

Sin duda, este caso puso de manifiesto tensiones críticas entre la innovación tecnológica y las garantías procesales, especialmente respecto a la opacidad algorítmica, el uso de datos protegidos como el género, y la aplicación de estadísticas grupales para evaluaciones individuales; representa, pues, un hito en la evolución jurisprudencial sobre el uso de la IA en el ámbito penal y genera importantes debates sobre transparencia, equidad y responsabilidad en los sistemas automatizados de cara a la toma de decisiones judiciales.

Así, pues, la vigilancia masiva y sus implicaciones para la privacidad son preocupaciones adicionales, pues la creciente integración de fuentes de datos diversas (con la inclusión de redes sociales, datos biométricos y sensores urbanos) para alimentar estos sistemas plantea serias preocupaciones sobre la vigilancia desproporcionada y el derecho a la privacidad. Asimismo, surge la cuestión de si una alerta generada algorítmicamente constituye “sospecha razonable” o “indicio suficiente” para justificar intervenciones como identificaciones, registros o detenciones preventivas.

En respuesta a estos desafíos, diversas jurisdicciones y sistemas penales han comenzado a desarrollar marcos regulatorios específicos. Algunas ciudades como San Francisco y Oakland han prohibido completamente el uso de análisis predictivo por sus fuerzas policiales, al estimar que los riesgos superan potenciales beneficios. Nueva York, a su vez, introdujo la “POST Act” (*Public Oversight of Surveillance Technology*), que obliga a divulgar públicamente qué tecnologías de vigilancia utiliza la policía y bajo qué protocolos. Dinamarca, para citar a un país europeo, estableció un comité ético independiente que evalúa obligatoriamente cualquier introducción de sistemas predictivos en contextos policiales; incluso, el proyecto de Reglamento sobre IA de la Unión Europea clasifica los sistemas predictivos criminales como de “alto riesgo”, e impone requisitos técnicos estrictos de transparencia, robustez y supervisión humana.

El análisis predictivo, pues, ilustra perfectamente la tensión inherente a la utilización de la IA en el ámbito penal: mientras promete optimizar recursos escasos y

potencialmente prevenir la victimización, simultáneamente desafía principios fundamentales como la presunción de inocencia, la no discriminación y la privacidad. Su adecuada regulación requiere, entonces, un enfoque interdisciplinario que combine la comprensión técnica con un sólido fundamento en los principios jurídico-penales y los derechos fundamentales.

c. La evidencia digital

La multiplicación de dispositivos digitales y el crecimiento de datos han cambiado radicalmente la naturaleza de la evidencia en investigaciones criminales. Actualmente, casi toda investigación penal compleja incluye elementos digitales: desde mensajes electrónicos hasta datos de ubicación, operaciones financieras o información técnica de archivos. Este nuevo escenario, llamado *forensic big data*, ha creado la necesidad de desarrollar herramientas de inteligencia artificial diseñadas específicamente para procesar, analizar e interpretar grandes volúmenes de información digital que superan ampliamente las capacidades humanas.

El mundo de la evidencia digital actual incluye diferentes categorías y cada una requiere métodos de análisis específicos. Las comunicaciones electrónicas comprenden correos, mensajería instantánea, publicaciones en redes sociales y mensajes cifrados; los sistemas de IA aplicados a estas comunicaciones incluyen programas que procesan el lenguaje natural para identificar contenido importante y patrones lingüísticos, modelos que mapean redes de comunicación entre personas, sistemas que detectan comunicaciones inusuales, y herramientas que reconstruyen el orden de los acontecimientos.

Los datos de teléfonos móviles y dispositivos conectados (IoT) incluyen información extraída de celulares, dispositivos vestibles (dispositivos electrónicos que se llevan sobre el cuerpo), asistentes virtuales y aparatos domésticos inteligentes. Las herramientas de IA para estos datos incluyen programas para recuperar información borrada, sistemas que relacionan datos de diferentes dispositivos, análisis de patrones de uso para verificación, y procesamiento de registros de actividad. Por su parte, los datos financieros incluyen registros bancarios, criptomonedas y transacciones comerciales.

A su turno, las herramientas especializadas incluyen programas que detectan fraudes mediante aprendizaje automático, sistemas que analizan movimientos financieros sospechosos, modelos que identifican patrones de lavado de dinero, y herramientas para seguir el rastro de criptoactivos; también, los datos multimedia engloban imágenes, audio y video potencialmente relevantes. Las aplicaciones de IA incluyen reconocimiento facial y de objetos en material visual, verificación de autenticidad para detectar manipulaciones (*deepfakes*), análisis forense de audio para identificación de personas hablantes, y procesamiento automatizado de grabaciones de videovigilancia.

Los rastros digitales geoespaciales y temporales incluyen datos de geolocalización, registros de acceso y metadatos con marcas temporales; a su vez, las herramientas relevantes comprenden algoritmos de reconstrucción de trayectorias a partir de datos fragmentarios, sistemas de correlación espacio-temporal entre múltiples fuentes, y modelos de detección de inconsistencias para identificar manipulaciones.

Por eso, no es de extrañar que el uso de la inteligencia artificial haya revolucionado fundamentalmente las capacidades investigativas en el ámbito digital. Los sistemas modernos pueden analizar terabytes de información en tiempos reducidos, permitiendo -por ejemplo- que el FBI procesara 634.000 archivos incautados a Ghislaine Maxwell en menos de 24 horas (Caso *United States v. Maxwell*, 2021). Agréguese: Los algoritmos de aprendizaje automático pueden descubrir correlaciones sutiles indetectables mediante análisis humano; el sistema PALANTIR, utilizado por agencias como Europol, ha demostrado capacidad para conectar fragmentos aparentemente inconexos de información para identificar redes criminales complejas.

Las técnicas de recuperación basadas en IA permiten restaurar y analizar información deliberadamente borrada. El sistema *CELLEBRITE UFED*, para citar un caso, ha posibilitado extraer mensajes eliminados que resultaron cruciales en casos de alto perfil como el procesamiento de Patrick Frazee por homicidio, en 2019. Además, los algoritmos especializados pueden detectar manipulaciones en material audiovisual, algo crucial en una era en la cual las tecnologías de *deepfake* amenazan la fiabilidad probatoria; también, la combinación de datos históricos permite reconstruir eventos previos e inferir probables secuencias de acciones, facilitando la generación de hipótesis investigativas verificables.

La introducción de estos sistemas genera complejos desafíos procesales. Surge la cuestión de qué estándares deben cumplir los análisis basados en IA para ser admitidos como evidencia válida; con tales miras, jurisdicciones como las de Estados Unidos han comenzado a desarrollar criterios específicos como extensión de los estándares *Daubert/Frye* –esto es, criterios legales fundamentales utilizados en el sistema judicial para determinar la admisibilidad de evidencia científica y técnica en los tribunales– para evaluación de evidencia científica. La volatilidad inherente a la evidencia digital exige protocolos robustos que garanticen su integridad desde la recolección hasta la presentación judicial. Incluso, la doctrina especializada ha desarrollado el concepto de “continuidad de la evidencia digital” para adaptar el principio tradicional de cadena de custodia al entorno virtual.

Como es obvio, la complejidad técnica supone un obstáculo para el efectivo ejercicio del derecho de defensa, potencialmente vulnerando el principio de igualdad de armas. Por ejemplo, el Caso “*United States v. Budziak*” (2011) estableció que la defensa tiene derecho a acceder no solo a los resultados de análisis forenses digitales, sino también a los *software* y metodologías utilizados. Además, existe riesgo de sobrevaloración judicial de los resultados generados por sistemas de IA, percibidos frecuentemente como inherentemente objetivos y fiables. Sin embargo,

estos modelos presentan tasas de error específicas que deben ser comprendidas y consideradas.

También, la capacidad de procesar cantidades enormes de datos incentiva recolecciones excesivamente amplias. A este respecto, recuérdese que el Tribunal Europeo de Derechos Humanos, en el Caso “Big Brother Watch v. UK” (2021), estableció que la vigilancia masiva requiere salvaguardas proporcionadas, incluyendo supervisión independiente y limitaciones claras. La naturaleza transnacional de la evidencia digital, pues, frecuentemente almacenada en servidores ubicados en múltiples jurisdicciones, genera complejos problemas de derecho aplicable y cooperación internacional.

Para abordar estos desafíos, diversas jurisdicciones han comenzado a desarrollar marcos normativos específicos. Así, Organizaciones como SWGDE (*Scientific Working Group on Digital Evidence*) y ENFSI (*European Network of Forensic Science Institutes*) han desarrollado protocolos estandarizados para garantizar la fiabilidad de los análisis forenses digitales, incluyendo metodologías de validación para sistemas de IA. Incluso, jurisdicciones pioneras como las de los Países Bajos requieren divulgación completa de las metodologías utilizadas por sistemas de IA que generan evidencia en procesos penales, incluyendo tasas de error conocidas y limitaciones técnicas. La combinación de análisis automatizado con la verificación humana independiente emerge como estándar para mitigar riesgos de falsos positivos. Programas como el *Federal Judicial Center* en EE.UU. y la *European Judicial Training Network* han desarrollado módulos específicos para capacitar a jueces y fiscales en evaluación crítica de evidencia digital procesada mediante IA. Asimismo, se exige documentación exhaustiva de cada paso analítico realizado por sistemas automatizados, permitiendo la reconstrucción y la verificación retrospectiva.

El procesamiento digital de evidencia mediante IA representa, de forma simultánea, una oportunidad sin precedentes para mejorar las capacidades investigativas y un desafío fundamental para los principios procesales tradicionales. Por eso, su adecuada regulación requiere un equilibrio delicado entre innovación tecnológica y preservación de las garantías fundamentales, particularmente el derecho de defensa y la presunción de inocencia. La capacidad del sistema jurídico para adaptarse a esta realidad tecnológica determinará, pues, en gran medida la legitimidad del proceso penal de la era digital para no caer en manos de un expansionismo penal exacerbado que pisotee todas las garantías y arrase con un derecho penal mínimo como el propuesto por Ferrajoli (1995).

d. La reconstrucción objetiva de los hechos

El establecimiento preciso y objetivo de hechos delictivos constituye uno de los desafíos centrales en la investigación criminal; tradicionalmente dependiente de testimonios, evidencia física y análisis forense convencional, este proceso se encuentra en plena transformación gracias a la aplicación de sistemas avanzados de

inteligencia artificial. Estas tecnologías revolucionan la capacidad de las autoridades para recrear secuencias de eventos, verificar hipótesis investigativas y comprender dinámicas complejas en escenarios criminales, introduciendo simultáneamente nuevas metodologías forenses y complejas cuestiones jurídicas, para estudiar el lugar del hecho o el escenario (físico o virtual) del crimen.

Los sistemas contemporáneos de reconstrucción asistida por IA abarcan múltiples tecnologías especializadas. De hecho, la reconstrucción visual mediante *computer vision* incluye la fotogrametría forense automatizada, con algoritmos que generan modelos tridimensionales precisos de escenas del crimen a partir de fotografías convencionales; a su vez, sistemas como FARO Reality permiten recrear espacialmente escenarios criminales con precisión milimétrica, facilitando análisis de trayectorias balísticas, campos visuales de testigos y posiciones relativas. También incluye el análisis de patrones de salpicaduras de sangre mediante algoritmos que analizan distribuciones de fluidos biológicos para determinar mecánicas de impacto, alturas, ángulos y secuencias temporales.

El sistema HEMOSPAT, por su parte, con capacidades de *machine learning*, puede reconstruir dinámicas de agresiones con armas blancas o contundentes a partir de patrones de dispersión hemática. De igual forma, la reconstrucción facial forense utiliza sistemas que generan aproximaciones faciales a partir de restos óseos, utilizando modelos estadísticos de tejidos blandos entrenados con miles de CT *scans*; incluso el sistema TEMS (*Tissue Estimation for Morphometric Software*) demuestra una precisión muy superior a métodos tradicionales en materia de identificación de personas desaparecidas.

Además, el análisis algorítmico de trayectorias y movimientos incluye la reconstrucción de accidentes y colisiones mediante simulaciones físicas basadas en *machine learning* que integran datos de sensores, marcas de frenado, deformaciones vehiculares y posiciones finales para reconstruir dinámicas de siniestros viales. Incluso, en materia de balística, sistemas como PC-CRASH con módulos de IA pueden determinar velocidades, puntos de impacto y secuencias con precisión superior a métodos convencionales; el análisis de trayectorias balísticas interiores utiliza algoritmos que simulan comportamientos de proyectiles dentro de espacios cerrados, considerando rebotes, fragmentaciones y pérdidas energéticas. Incluso, el sistema ABIS (*Advanced Ballistic Intelligence System*) puede reconstruir intercambios de disparos complejos, incluyendo la identificación de las posiciones de los tiradores involucrados.

De la misma manera, la recreación de secuencias temporales se realiza mediante la correlación multievento, con sistemas que integran datos temporales fragmentarios (transacciones, comunicaciones, registros de acceso, geolocalización) para reconstruir líneas temporales coherentes. En este contexto plataformas como MALTEGO, con capacidades de IA, permiten visualizar y analizar secuencias complejas de eventos relacionados con actividades criminales; incluso la reconstrucción de metadatos degradados utiliza algoritmos que recuperan y correlacionan información temporal

parcialmente perdida o manipulada en archivos digitales, algo que es crucial para establecer cronologías en los casos de fraude documental o encubrimiento.

La simulación cognitiva y comportamental incluye el modelado de conductas criminales con sistemas que simulan patrones basados en perfiles criminológicos, para generar hipótesis sobre secuencias probables de acción. El FBI ha desarrollado el sistema VICAP (*Violent Criminal Apprehension Program*) potenciado con IA para identificar patrones consistentes con comportamientos delictivos conocidos; además, el análisis en torno a la viabilidad de hipótesis utiliza algoritmos que evalúan la plausibilidad física, temporal y psicológica de diferentes escenarios reconstructivos propuestos, identificando inconsistencias lógicas o físicas.

Como es de suponer, la introducción de estas tecnologías ha producido resultados significativos en investigaciones complejas; en el caso del atentado terrorista de Manchester Arena (2017), que generó 22 muertos y 800 heridos en dicha ciudad del Reino Unido, sistemas de reconstrucción asistida por IA permitieron integrar más de 14.000 horas de material de videovigilancia para recrear los movimientos del perpetrador durante las semanas previas, identificando patrones de reconocimiento y planificación previamente inadvertidos.

Asimismo, la investigación del derribo del vuelo MH17 sobre Ucrania el 17 de julio de 2014 (el avión de la Malaysia Airlines cubría la ruta Ámsterdam-Kuala Lumpur y estalló dejando 298 víctimas), utilizó algoritmos de reconstrucción balística avanzada para determinar la trayectoria exacta del misil BUK, su punto de lanzamiento y logró vincularlo con unidades militares específicas; la evidencia crucial fue presentada ante la Corte Internacional Penal y su Fiscalía. Añádase que el Departamento de Policía de Los Ángeles introdujo el sistema ECHO (*Event Chronology and Hypothesis Optimization* u “Optimización de Cronología de Eventos e Hipótesis”) para casos de homicidio complejos, reportando un incremento del 23% en la tasa de resolución de los mismos mediante la identificación de inconsistencias temporales imperceptibles en coartadas aparentemente sólidas.

Desde luego, la utilización de estas tecnologías genera importantes desafíos en el ámbito jurídico. Obsérvese: Las reconstrucciones basadas en IA plantean interrogantes sobre su aceptabilidad como evidencia científica; en jurisdicciones o sistemas penales que aplican los estándares Daubert (como los EE.UU.) o similares, surge la cuestión de cómo verificar la validez metodológica de sistemas cuyo funcionamiento interno puede resultar opaco incluso para sus desarrolladores. El caso “*State v. Loomis*” (Wisconsin, 2016), ya mencionado, estableció un precedente parcial en esta materia al permitir el uso de sistemas algorítmicos, pero imponiéndose restricciones significativas sobre su peso y valor probatorio.

Estos sistemas pueden reforzar inadvertidamente hipótesis investigativas iniciales, incorporando sesgos confirmatorios automatizados, lo cual es explicable cuando se tiene en cuenta que la IA está compuesta de algoritmos, y, por tanto, está sujeta

a que ocurran sesgos algorítmicos en virtud de los cuales los participantes en la vida del *software* “filtran” sus influencias en los agentes inteligentes por medio del código, el diseño según los requisitos del sistema, la implementación, el mantenimiento o la retroalimentación (*feedback*) que recibe y de la cual se sirve el sistema de autoaprendizaje (Davies & Douglas, 2020); y en esos ciclos de vida de la IA intervienen personas que tienen comportamientos racistas, sexistas, u otros sesgos, que se transmiten a los algoritmos que utilizan los agentes y su racionalidad estará condicionada a reflejar los sesgos que tienen los participantes de su ciclo de vida. Incluso, cuando intervienen analistas forenses ellos tienden a interpretar resultados ambiguos de sistemas de IA de forma consistente con información contextual previa; por eso se habla del fenómeno denominado como “cascada de sesgo cognitivo-tecnológico”.

Asimismo, la complejidad técnica de reconstrucciones avanzadas puede dificultar su adecuada comprensión por parte de juzgadores legos, potencialmente devaluando su utilidad procesal o, contrariamente, generando una sobrevaloración acrítica de su aparente objetividad tecnológica. Incluso, la naturaleza de la propiedad de muchos sistemas plantea problemas para el ejercicio efectivo del derecho de defensa, particularmente cuando algoritmos y metodologías específicas están protegidos como secretos comerciales. El Caso “People v. Superior Court (Chubbs)”, en California, es paradigmático porque él destacó esta tensión entre la propiedad intelectual y el derecho de defensa respecto al *software* forense. En este contexto, obsérvese, existe el riesgo de presentar como certezas lo que en realidad son reconstrucciones probabilísticas basadas en datos incompletos, debido a la capacidad de estos sistemas para generar visualizaciones detalladas y aparentemente exactas.

Para abordar estos importantes desafíos, diversas jurisdicciones y sistemas penales han comenzado a desarrollar marcos regulatorios específicos. Países como Holanda y Australia desarrollan protocolos específicos para validar científicamente sistemas de reconstrucción forense automatizada antes de su utilización judicial, para lo cual se requieren estudios de fiabilidad, precisión y tasas de error conocidas. Además, sistemas judiciales avanzados como el de Canadá establecen obligaciones de divulgación completa de parámetros, asunciones y limitaciones de los sistemas de IA utilizados en reconstrucciones forenses, incluyendo la indicación de los márgenes de error y confianza estadística.

Emerge, entonces, como buena práctica la exigencia de una constatación independiente por medio de métodos alternativos, para evitar decisiones basadas solo en reconstrucciones algorítmicas. Incluso, se han desarrollado programas de capacitación para jueces y fiscales en interpretación crítica de reconstrucciones forenses basadas en IA, como los confeccionados por el *National Judicial College* en EE.UU. y la Academia de Derecho Europeo (ERA); también, se han establecido directrices para la presentación de reconstrucciones en juicio que exigen la contextualización explícita de las limitaciones, las alternativas plausibles y el grado de certeza, evitando presentaciones que puedan ser engañosas por su aparente precisión visual.

VI. CONCLUSIONES

La incidencia de la inteligencia artificial en las construcciones teóricas del derecho penal, parte general, no constituye un fenómeno periférico, sino una transformación que afecta al núcleo mismo de esta disciplina; incluso a la parte especial mediante la necesaria reconfiguración de diversos delitos. Las categorías conceptuales tradicionales –acción, tipicidad, antijuridicidad, culpabilidad, bien jurídico, pena, etc.– experimentan cambios profundos que requieren una adaptación constructiva de la dogmática penal.

Por eso, lejos de adoptar posiciones tecnofóbicas o tecnofílicas extremas, la ciencia penal contemporánea debe desarrollar un enfoque equilibrado que permita, por un lado, preservar garantías y principios fundamentales del derecho penal y, por otro, acomodar sus categorías conceptuales para responder adecuadamente a los desafíos que plantea la IA. Esta adaptación exige un diálogo interdisciplinar fructífero entre juristas, científicos computacionales, filósofos de la tecnología y éticos, que permita desarrollar construcciones dogmáticas coherentes con la realidad tecnológica actual y futura. Solo mediante este esfuerzo colaborativo será posible preservar la función del derecho penal como instrumento de control social legítimo en la era digital.

Esta ampliación de perspectiva confirma que la transformación del derecho penal ante la IA no constituye un fenómeno aislado o sectorial, sino una reconfiguración sistémica que afecta transversalmente a la totalidad de categorías e instituciones dogmáticas, requiriendo un esfuerzo reconstructivo que mantenga la coherencia interna del sistema mientras incorpora las particularidades de los entornos tecnológicamente mediados.

De la misma manera, la integración de la IA en el sistema procesal penal representa simultáneamente una oportunidad para mejorar la eficacia investigativa y judicial, y un desafío fundamental para principios constitutivos del Estado de Derecho como la presunción de inocencia, la igualdad procesal, la independencia judicial y el derecho de defensa. Los sistemas de evaluación de riesgos y apoyo a la decisión judicial deben diseñarse y regularse para funcionar como herramientas complementarias que amplíen la información disponible para los operadores jurídicos, nunca como sustitutos del criterio judicial fundamentado.

En el ámbito de la investigación criminal, las herramientas de reconocimiento facial, análisis predictivo, procesamiento de evidencia digital y reconstrucción de hechos transforman radicalmente las capacidades investigativas pero plantean interrogantes serios sobre la privacidad, no discriminación y debido proceso. El desarrollo de marcos regulatorios específicos, como los que ya han comenzado a implementarse en diversas jurisdicciones, resulta indispensable para aprovechar el potencial de estas tecnologías sin comprometer derechos fundamentales.

El análisis comparativo internacional revela una tensión compartida en diversos sistemas jurídicos entre la adopción acelerada de estas tecnologías y la insuficiencia de los marcos normativos existentes para regularlas adecuadamente. Las lecciones aprendidas de la experiencia acumulada sugieren la necesidad de principios regulatorios comunes como la transparencia algorítmica, la supervisión independiente, la subsidiariedad tecnológica y la salvaguarda efectiva de derechos procesales.

La revolución de la IA en el sistema penal no constituye, pues, una mera actualización tecnológica, sino una transformación paradigmática que obliga a repensar categorías fundamentales tanto en el plano sustantivo como procesal. El reto para legisladores, tribunales, académicos y funcionarios judiciales consiste en desarrollar un enfoque que, sin caer en el rechazo tecno fóbico ni en la aceptación acrítica, permita integrar estas tecnologías de forma que fortalezcan, en lugar de debilitar, los principios fundamentales del Estado de Derecho.

La transparencia en el uso de algoritmos es esencial para garantizar el derecho de defensa. Sin embargo, muchos sistemas de IA están protegidos como secretos comerciales, lo que dificulta que los abogados y los jueces entiendan cómo se toman las decisiones. Para solucionar esto, se podrían implementar leyes que exijan a las empresas compartir los códigos fuente de los algoritmos utilizados en procesos judiciales, siempre que no comprometan la propiedad intelectual (Burchard, 2019).

VII. REFERENCIAS

- Agustina Sanllehi, J. R. (2022). La inmersión del Derecho penal en la era de la "postmodernidad tecnológica": retos en la adaptación del sistema jurídico-penal ante un cambio de paradigma. En José Antonio Pérez Sanjuán & Francisco Javier Sanjuán Andrés: *Cuadernos Digitales. Derecho y Nuevas Tecnologías* (pp. 219-249). Aranzadi. «»
- Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *Machine bias*. ProPublica. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- Barrio Andrés, M. et al. (2019). *Derecho de los robots* (2.^a ed.). Wolters Kluwer.
- Bell, K., Hong, J., McKeown, N. & Voss, C. (2021). The Recon Approach: a New Direction for Machine Learning in Criminal Law. *Berkeley Technology Law Journal* 37, 102-139. <https://ssrn.com/abstract=3834710>
- Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 81, 77-91.

- Burchard, Ch. (2019). Künstliche Intelligenz als Ende des Strafrechts? Zur algorithmischen Transformation der Gesellschaft. *Jahrbuch für Recht und Ethik. Strafrecht und Rechtsphilosophie Zugleich Gedächtnisschrift für Joachim Hruschka*, 27, 527-556.
- Calaza López, S. & Llorente Sánchez-Arjona, M. (Dirs.) et al. (2022). *Inteligencia artificial legal y administración de Justicia*. Thomson Reuters-Aranzadi.
- Campbell, M. Hoane Jr., A. J., & Feng-hsiung, H. (2002). Deep Blue. *Artificial Intelligence* 134 (2002) 57–83. <https://www.mimuw.edu.pl/~ewama/zsi/deepBlue.pdf>
- Consejo Superior de la Judicatura (2024, diciembre 16). Acuerdo PCSJA24-12243. Por el cual se adoptan lineamientos para el uso y aprovechamiento respetuoso, responsable, seguro y ético de la Inteligencia Artificial en la Rama Judicial. https://actosadministrativos.ramajudicial.gov.co/GetFile.ashx?url=%7E%2FApp_Data%2FUpload%2FPCSJA24-12243.pdf
- Corte Constitucional (2020, febrero 10). Sentencia T-043 [Exp. T-7.461.559]. Magistrado Ponente; José Fernando Reyes Cuartas. [Rama judicial, Colombia].
- Corte Constitucional (2024, agosto 4). Sentencia T-323 [Exp. T-9.301.656] Magistrado Ponente: Juan Carlos Cortés González. [Rama judicial, Colombia].
- Corte Constitucional (2025, febrero 26). Sentencia T-067 [Exp. T-8.202.533] Magistrado Ponente: Natalia Ángel Cabo. [Rama judicial, Colombia].
- Davies, B., & Douglas, T. (17 de Noviembre de 2020). *Learning to Discriminate: The Perfect Proxy Problem in Artificially Intelligent Criminal Sentencing*. Obtenido de PhilArchive: <https://philarchive.org/archive/DAVLTD-3v1>
- De Hoyos Sancho, M. (2020): El libro Blanco sobre la inteligencia artificial de la Comisión Europea: Reflexiones desde las garantías del proceso penal como sector de riesgo», *Revista Española de Derecho Europeo*, 76, 9-43.
- Desmarais, S. L., & Singh, J. P. (2013). Risk assessment instruments validated and implemented in correctional settings in the United States. Council of State Governments Justice Center. <https://csgjusticecenter.org/wp-content/uploads/2020/02/Risk-Assessment-Instruments-Validated-and-Implemented-in-Correctional-Settings-in-the-United-States.pdf>
- Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1), <https://www.science.org/doi/10.1126/sciadv.aao5580>
- Ferrajoli, L. (1995). *Derecho y Razón. Teoría del garantismo penal* [trad. P. Andrés Ibáñez et al.]. Trotta.
- Forero-Corba, W. & Negre Bennasar, F. (2024). Técnicas y aplicaciones del Machine Learning e Inteligencia Artificial en educación: una revisión sistemática. *RIED-*

- Revista Iberoamericana de Educación a Distancia*, 27(1), 1-34. <https://revistas.uned.es/index.php/ried/article/view/37491>
- Glaser, S. (2024). Künstliche Intelligenz im Strafrecht. – *Zeitschrift für Polizeiwissenschaft und polizeiliche Praxis* 1, 10-21.
- Gómez Colomer, J. L. (2023). *El juez robot. La independencia judicial en peligro*. Tirant lo Blanch.
- Green, B., & Chen, Y. (2021). Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), 1-33. <https://arxiv.org/abs/2012.05370>
- 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://academic.oup.com/qje/article-abstract/133/1/237/4095198>
- Ibold, V. (2024). *Künstliche Intelligenz und Strafrecht. Zur strafrechtlichen Produktverantwortung in der Innovationsgesellschaft*. Nomos Verlag.
- Lum, K., & Isaac, W. (2016). To predict and serve? *Significance*, 13(5), 14-19. <https://academic.oup.com/jrssig/article/13/5/14/7029190m>
- Martinez, R. (2019). Artificial Intelligence: Distinguishing Between Types & Definitions. *Nevada Law Review* 19(3), 1015-1042. <https://scholars.law.unlv.edu/cgi/viewcontent.cgi?article=1799&context=nlj>
- Martínez Garay, L. (2018): «Peligrosidad, algoritmos y due process: El caso State v. Loomis», *Revista de Derecho Penal y Criminología*, 20, 482-502.
- McCarthy, J. (2007). What is artificial intelligence? Stanford University. <http://www-formal.stanford.edu/jmc/whatisai/whatisai.html>
- McCarthy, J.; Minsky, M.; Rochester, N. & Shannon, C. (1955). A Proposal for The Dartmouth Summer Research Project on Artificial Intelligence. Stanford University, Computer Science Department. EEUU. <http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf>
- Mohler, G. O., Short, M. B., Malinowski, S., Johnson, M., Tita, G. E., Bertozzi, A. L., & Brantingham, P. J. (2015). Randomized controlled field trials of predictive policing. *Journal of the American Statistical Association*, 110(512), 1399-1411. <https://mshort9.math.gatech.edu/papers/MohlerEtAl-2015-JASA-Predictive-InPress.pdf>
- Muñoz Conde, F. & García Arán, M. (2022). *Derecho Penal, Parte General* (11.ª ed.). Tirant lo Blanch.
- Nieva Fenoll, J. (2022). *Inteligencia artificial y proceso judicial*. Marcial Pons.

- Papini, G. (1973). *Gog. El Libro Negro*. Círculo de Lectores.
- Planchadell Gargallo, A. (2024). Inteligencia artificial y proceso penal; especial referencia las medidas cautelares. *Teorder* 37, 176-195.
- Posada Maya, R. (2017). El cibercrimen y sus efectos en la teoría de la tipicidad: de una realidad física a una realidad virtual, *Nuevo Foro Penal*, 13(88), 72-112.
- Rademacher, T. (2020). Künstliche Intelligenz und neue Verantwortungsarchitektur. M. Eifert (Edit.), *Digitale Disruption und Recht. Workshop zu Ehren des 80. Geburtstags von Wolfgang Hoffmann-Riem* (pp. 45-72). Nomos Verlag.
- Royal United Services Institute for Defense and Security Studies. *Annual Report of Trustees* (2023, marzo 31). *Charity N° 210639*. https://static.rusi.org/rusi_financial_statements_-_31_march_2020_0.pdf
- Russell, S., & Norvig, P. (2008). *Inteligencia Artificial. Un enfoque moderno*. (2.ª ed.). Pearson Educación S. A.
- Russell, S., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson.
- Staffler, L. & Jany, O. (2020). Künstliche Intelligenz und Strafrechtspflege – eine Orientierung. *Zeitschrift für Internationale Strafrechtsdogmatik (ZIS)* 4, 164-177.
- Seher, G. (2016). Intelligente Agenten als „Personen“ im Strafrecht? En Sabine Gless/ Kurt Seelmann (Dirs.): *Intelligente Agenten und das Recht* (pp. 45-60). Nomos Verlag.
- Silva Sánchez, J. M. (2025). *Derecho Penal, Parte General*. Civitas.
- Stevenson, M. T. (2018). Assessing risk assessment in action. *Minnesota Law Review*, 103, 303-384. https://www.minnesotalawreview.org/wp-content/uploads/2019/01/13Stevenson_MLR.pdf
- Stevenson, M. T., & Doleac, J. L. (2021). Algorithmic risk assessment in the hands of humans. IZA Discussion Paper No. 14312. <https://ssrn.com/abstract=3835131>
- Turing, A. (1950). Computing Machinery ad Intelligence. *Mind. A Quarterly Review of Psychology and Philosophy* 59 (236), 433-460. <https://phil415.pbworks.com/f/TuringComputing.pdf>
- Unión Europea (2016). Reglamento (UE) 2016/679 Parlamento Europeo y del Consejo de 27 de abril de 2016. relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE (Reglamento general de protección de datos). <https://www.boe.es/buscar/doc.php?id=DOUE-L-2016-80807>
- Unión Europea (2024). Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas

- en materia de Inteligencia Artificial y por el que se modifican los Reglamentos (CE) n.º 300/2008, (UE) n.º 167/2013, (UE) n.º 168/2013, (UE) 2018/858, (UE) 2018/1139 y (UE) 2019/2144 y las Directivas 2014/90/UE, (UE) 2016/797 y (UE) 2020/1828 (Reglamento de Inteligencia Artificial). <https://www.boe.es/doue/2024/1689/L00001-00144.pdf>
- Valls Prieto, J. (2023). La inteligencia artificial frente a la colectividad. Una protección de los bienes jurídicos colectivos frente al auge de los sistemas inteligentes. *Estudios Penales y Criminológicos* 44, num. Ext., 1-26.
- VanNostrand, M., & Rose, K. (2009). Pretrial risk assessment in Virginia. Virginia Department of Criminal Justice Services. <https://www.dcjs.virginia.gov/sites/dcjs.virginia.gov/files/publications/corrections/virginia-pretrial-risk-assessment-report.pdf>
- Velásquez, F. (2025). *Derecho Penal, Parte General* (7.^a ed.). Tirant lo Blanch.