REVISTA
DE ESTUDIOS DE LA JUSTICIA

NÚM. 42 (2025) • PÁGS. 153-183 • DOI 10.5354/0718-4735.2025.77061
RECIBIDO: 16/12/2024 • APROBADO: 26/6/2025 • PUBLICADO: 31/7/2025

GENERAL

Responsabilidad penal en la era de la inteligencia artificial: De la agencia humana a la autonomía de la *machina sapiens*

Criminal liability in the age of artificial intelligence: From human agency to machina sapiens autonomy

Pablo Aguilar Campos y Víctor Alé Martínez
Investigadores independientes, Chile

RESUMEN En el marco del desarrollo acelerado de tecnologías basadas en inteligencia artificial (IA), este artículo examina los desafíos que plantea la atribución de responsabilidad penal tanto a los sistemas de IA como a los agentes humanos que los diseñan, programan o utilizan. Con este propósito, se revisarán los principales modelos de imputación penal, diferenciando entre la responsabilidad que podría recaer en desarrolladores, fabricantes y programadores, y aquella atribuible a quienes empleen estas tecnologías como instrumentos para la perpetración de delitos. Posteriormente, se explora la hipótesis de que ciertas formas avanzadas de IA, particularmente dotadas de alta autonomía funcional, puedan, en un futuro, ser consideradas penalmente responsables. Finalmente, se expone sobre la naturaleza y límites de las sanciones aplicables en escenarios de agencia artificial no humana.

PALABRAS CLAVE Culpabilidad, persona jurídica, agencia artificial, autoría, castigo penal.

ABSTRACT In the context of the accelerated development of artificial intelligence (AI) based technologies, this article examines the challenges posed by the attribution of criminal responsibility both to AI systems and to the human agents who design, program, or use them. To this end, it reviews the main models of criminal imputation related to AI, distinguishing between the liability that may fall on developers, manufacturers, and programmers, and that attributable to individuals who employ these technologies as instruments for the commission of criminal offenses. The article then explores the hypothesis that certain advanced forms of AI, particularly those endowed with a high degree of functional autonomy, could, in the future, be considered criminally responsible. Finally, it reflects on the nature and limits of the sanctions applicable in scenarios involving non-human artificial agency.

KEYWORDS Culpability, legal entity, artificial agency, perpetration, criminal punishment.

Introducción

La inteligencia artificial (IA) ha transformado de manera profunda la vida en sociedad del ser humano, posiblemente a una escala incluso superior a la de la Revolución Industrial (Blanco, 2019: 63),[1] constituyéndose en una realidad constatable (Araya, 2020: 258). Hoy en día, la IA está presente en nuestra vida cotidiana a través de aplicaciones de uso masivo —ChatGPT, Google Gemini o plataformas de Meta—, así como en sistemas altamente especializados, tales como programas (software) de pilotaje automático de aeronaves, cirugías asistidas por robots, vehículos terrestres autopilotados, negociaciones algorítmicas de alta frecuencia, sistemas de control y monitoreo del tráfico viario, entre otras (Navarro y Vidal, 2021: 262; Xavier, 2023: 4 y 5). Ante un cambio de tal magnitud, se prevén importantes beneficios para la comunidad en el corto y mediano plazo, pero también surgen grandes desafíos y riesgos criminógenos de diversa índole.

En el ámbito del derecho comparado existe amplio consenso en cuanto a que, en un futuro próximo, la IA podría alcanzar capacidades funcionalmente equiparables —e incluso, para algunos autores, exactamente iguales— a las del pensamiento, razonamiento e inteligencia humana (Fahim y Bajpai, 2020: 64). Esta evolución permitiría que los sistemas operen de manera autónoma, sin necesidad de intervención o instrucción humana directa, fenómeno que se ha conceptualizado como *Strong AI* o *machina sapiens* (Hallevy, 2013: 5). En efecto, esta tesis ha sido sostenida por académicos del derecho penal pertenecientes a sistemas jurídicos diversos, tales como Rusia (Kirpichnikov y otros, 2020: 1-3), Estados Unidos (Hallevy, 2013: 19-21), India (Fahim y Bajpai, 2020: 64-94) y España (Blanco, 2019: 63-80), entre otros. Incluso se ha llegado a plantear la posibilidad de reconocer a este tipo de IA una «personalidad legal» (Hildebrandt, 2020; Kurko, 2019).

En este marco, este artículo busca contribuir a esclarecer algunas de las principales controversias en torno a la atribución de responsabilidad penal vinculada a la inteligencia artificial, tanto para el agente humano como para la *Strong IA* o *machina sapiens*.

1. En este sentido, ha sido catalogada como la «Cuarta Revolución Industrial» (Sánchez y Toro-Valencia, 2021: 212; Barona, 2019: 2), la que estaría caracterizada porque «extrae su energía de la abundancia de datos combinada con potentes algoritmos y capacidad informática, expansiva a nivel mundial, con una rápida convergencia y el enorme impacto de los nuevos avances tecnológicos en los países, las economías, las sociedades, las relaciones internacionales y el medio ambiente, lo cual supone un cambio radical de enorme magnitud que repercute diversamente en las distintas partes de la sociedad, en función de sus objetivos, ubicación geográfica o contexto socioeconómico» (Morillas, 2023: 63).

Conceptualización de la inteligencia artificial y su vinculación con el derecho penal

Definición y características

No existe consenso respecto de cómo conceptualizar a la IA (Sánchez y Toro-Valencia, 2021: 213; Hernández, 2019: 794 y 795; Valls, 2022: 3), tanto en lo relativo a su contenido como a su extensión.[2] No obstante, Valls, examinando las diversas definiciones disponibles de la IA, identifica los siguientes elementos comunes: i) percepción del entorno (incluyendo la consideración de la complejidad del mundo real); ii) procesamiento de información (a través de la recopilación e interpretación de *inputs* en forma de datos); iii) toma de decisiones (que implica razonamiento y aprendizaje); iv) realización de tareas con ciertos niveles de autonomía; y v) el logro de objetivos específicos (Valls, 2022: 7).

Por su parte, el especialista en *AI Law* (una nueva rama del derecho en Estados Unidos), McCarl, ofrece una caracterización integral, sosteniendo que el software de la IA desempeña funciones asimilables con la mente humana, tales como la percepción, el reconocimiento de patrones, clasificar, razonar y procesar lenguaje, además de otras propiedades singulares: es frecuentemente autónoma (esto es, con capacidad de autogestión), está orientada a objetivos y posee una capacidad de automejora (McCarl, 2022: 95).

Por su parte, la Organización para la Cooperación y el Desarrollo Económico (OCDE) define la IA como un «sistema basado en una máquina que puede, para un conjunto determinado de objetivos definidos por el ser humano, hacer predicciones, recomendaciones o tomar decisiones que influyan en entornos reales o virtuales. Los sistemas de IA están diseñados para funcionar con distintos niveles de autonomía» (2022: 7).

Una definición particularmente destacada por su claridad y amplitud para fines jurídicos (Valls, 2022: 11 y 32) es la propuesta por el Grupo de Expertos en Inteligencia Artificial de la Comisión Europea, en los siguientes términos:

> Los sistemas de inteligencia artificial (IA) son sistemas de software (y posiblemente también hardware) diseñados por humanos que, dado un objetivo complejo, actúan en la dimensión física o digital percibiendo su entorno mediante la adquisición de datos, interpretando los datos recogidos estructurados o no estructurados, razonando sobre los conocimientos o procesando la información, derivados de estos datos y decidiendo la(s) mejor(es) acción(es) a tomar para alcanzar el objetivo dado.

2. De la Cuesta releva las dificultades de conceptualización asociadas al contenido (¿qué es la inteligencia?) y a la extensión (¿solo son inteligentes los humanos?) (2019: 52).

> Los sistemas de IA pueden utilizar reglas simbólicas o aprender un modelo numérico, y también pueden adaptar su comportamiento analizando cómo se ve afectado el entorno por sus acciones anteriores (2019: 6).

A lo anterior se suma la definición contenida en la Ley de Inteligencia Artificial (Reglamento [UE] 2024/1689), de 12 de junio de 2024, en cuyo artículo 3 consigna:

> «Sistema de IA»: Un sistema basado en una máquina que está diseñado para funcionar con distintos niveles de autonomía y que puede mostrar capacidad de adaptación tras el despliegue, y que, para objetivos explícitos o implícitos, infiere de la información de entrada que recibe la manera de generar resultados de salida, como predicciones, contenidos, recomendaciones o decisiones, que pueden influir en entornos físicos o virtuales.[3]

Pues bien, como se desprende de las definiciones expuestas, lo paradigmático de la IA radicaría en su aspecto relacional con la inteligencia humana como parámetro de comparación e imitación, desde el pensamiento y la actuación (Araya, 2020: 259). En tal sentido, la IA buscaría reproducir funcionalmente, mediante un proceso de simulación, el razonamiento humano y los procesos cognitivos, particularmente en lo relativo a la toma de decisiones, solución de problemas e, inclusive, capacidad de aprendizaje (Martínez, 2012: 828; Miró, 2018: 91 y 92). En este sentido, aunque la máquina no pueda replicar el pensamiento humano a nivel neurofisiológico, se considera que sería capaz de simular, funcionalmente, tal proceso de pensamiento (Martínez, 2012: 829).

Sin perjuicio de los matices de las definiciones propuestas, estimamos que un aspecto a considerar es que ciertas formas avanzadas de IA —como la denominada *fuerte*, que se tratará a continuación—, podrían, en un futuro próximo, operar sin necesidad de depender de fines previamente determinados por seres humanos, adquiriendo, por tanto, una capacidad de autodeterminación funcional. En consecuencia, puede sostenerse que la IA se caracteriza, en términos estructurales, por su aptitud para ejecutar procesos mediante instrucciones primigenias, con distintos grados de autonomía operativa sin supervisión humana constante, y exhibiendo un margen variable de imprevisibilidad en su actuar o comportamiento.

Sobre la clasificación funcional de la inteligencia artificial: Débil, fuerte y superinteligente

En doctrina, se ha propuesto una clasificación funcional tripartita de la IA: i) la *débil, estrecha* o *específica*, que posee un rango limitado de habilidades; ii) la *fuerte*, *general*

3. Disponible en https://tipg.link/gPxV. Próximamente, es esperable que el Estado de California también apruebe la primera ley de IA en los Estados Unidos. Disponible en https://tipg.link/gPxZ.

o *profunda*, que replica las capacidades cognitivas humanas; y iii) la *superinteligente*, que desarrollaría capacidades superiores a las humanas (del Rosal Blasco, 2023: 9).

La *IA débil* se caracteriza por estar diseñada para tareas concretas y delimitadas, actuando en función de datos y fines que le son proveídos y ordenados por el ser humano (Miró, 2025: 180; La Parra, 2021: 19).[4] Su operatividad, por tanto, se limita a objetivos predefinidos, ciñéndose a los *inputs* que recibe, sin exceder su programación original. Ejemplos de este tipo de tecnologías son los asistentes virtuales Siri o Alexa; softwares de reconocimiento facial, programas como Google Search, entre otros (del Rosal Blasco, 2023: 9). Esta tipología, compuesta por las también denominadas *dumb machines*, carece de capacidad de aprendizaje autónomo, de razonamiento complejo o de toma de decisiones fuera del marco delimitado por sus desarrolladores, asimilándose más bien a un modelo de funcionamiento matemático (Miró, 2025: 190).

La *IA fuerte* —por ahora, hipotética— es aquella que se comportaría de forma autónoma, pero con capacidades equivalentes a la acción humana (Miró, 2025: 180). Esta tipología buscaría replicar o equiparar la inteligencia y los comportamientos humanos, pero diferenciándose de una simple simulación a partir de habilidades de aprendizaje automático y la utilización de sus aptitudes con el objeto de resolver cualquier tipo de problema, sin necesidad de ser programada expresamente para ello (del Rosal Blasco, 2023: 9; La Parra, 2021: 20).

En esta línea, Kirpichnikov y sus colaboradores consideran que sería preferible utilizar la designación de *IA fuerte (Strong AI)* para referirse adecuadamente a una entidad como la *machina sapiens* (distinguiéndola, por ejemplo, de los «sistemas expertos» u otros «sistemas electrónicos débiles») (2020: 4). Esta visión es compartida por Kurki, para quien una IA fuerte «es una entidad que puede, en ámbitos relevantes, actuar como un ser humano», agregando que «tales IA tarde o temprano existirán» (2019: 177). Asimismo, según Hallevy, la conceptualización de la *IA fuerte* frecuentemente se asocia con «la habilidad de una máquina para imitar el comportamiento

4. El *machine learning* y el *deep learning*, en su actual configuración, son técnicas de procesamiento propias de la IA débil, y no se corresponden, estrictamente, con la IA general o fuerte, la cual sigue siendo una hipótesis futura en desarrollo (Miró, 2025: 180). Dicho esto, a pesar de sus aparentes notas comunes, se ha planteado que la *machine learning* (aprendizaje de las máquinas mientras se incorporan datos actualizados) (Miró, 2018: 91), no es equivalente a la IA *stricto sensu*, sino que constituye un subcampo o rama de esta (McCarl, 2022: 926; en idéntico sentido, Lior, 2020: 1057). A su vez, el *deep learning* —entendido como la imitación de una red neuronal humana, a través de capas individuales de conexión para realizar tareas específicas— sería una subcategoría dentro del *machine learning* (supervisado o no) (McCarl, 2022: 929). Se ha sostenido, además, que el *deep learning* más avanzado sigue siendo «una caja negra», en tanto no puede explicar actualmente cómo llega a decisiones o hace predicciones (McCarl, 2022: 941). Una distinción similar puede establecerse respecto de los diversos modelos de procesamiento de datos o información, a saber: *big data* (gestión de grandes volúmenes de datos) y *data mining* (encontrar patrones y resumir volúmenes de datos para la ulterior toma de decisiones) (Morán, 2021: 294 y 295).

inteligente» mediante «la simulación de los procesos cognitivos y el comportamiento humano». Sobre esta base, dicho autor plantea que la creación de una verdadera máquina pensante resultaría equivalente a la aparición de una *nueva especie de individuos en la tierra*, a saber, la *machina sapiens* (2013: 5).

La *IA superinteligente*, por su parte, estaría todavía lejos de existir. Esta no se limitaría a replicar o a comprender la inteligencia y el comportamiento humano, sino que, como afirma del Rosal Blasco, haría de las máquinas seres conscientes de sí mismos, superando la capacidad de la inteligencia y las habilidades humanas (2023: 9). En este contexto, Hallevy acuña el concepto *machina sapiens criminalis*, concebida como un subproducto de la *machina sapiens ideal* (2013: 18). Esta nueva clase de IA sería superior a las de *inteligencia limitada (dumb machines)*, pero inferior a la *machina sapiens ideal*, ubicándose «en algún lugar en el medio» entre ambas categorías. Sin embargo, dada su mayor proximidad al género de la *machina sapiens*, podría eventualmente ser considerada responsable en términos penales por sus actos (Hallevy, 2013: 21).

Esclarecida la distinción entre las clases de IA, se plantea que aquellas consideradas como fuertes y superinteligentes, podrían ser concebidas, al menos en términos teóricos, como agentes penalmente responsables, en la medida en que cumplan los requisitos de la imputación penal. En este sentido, Hallevy plantea que, «en tanto se satisfagan todos los requisitos relevantes del derecho penal, un individuo de nuevo tipo podría ser agregado al largo grupo de sujetos existentes para el derecho penal, en adición a los humanos y las corporaciones. Estos sujetos pueden ser referidos como *machina sapiens criminalis*» (2013: 21).[5]

A la luz de lo anterior, y sin perjuicio del carácter todavía limitado de estos sistemas, resulta pertinente examinar los modos en que la IA interactúa con el derecho penal, ya sea como herramienta auxiliar de los operadores, o como factor de riesgo en la producción de afectaciones a bienes jurídicos de relevancia.

Vinculación de la inteligencia artificial con el derecho penal

La IA tiene un impacto en diversas ramas del derecho, tales como las áreas de propiedad intelectual (derechos de autor de IA); civil (personalidad jurídica); laboral (selección y ejecución de empleos); en materia de privacidad (datos personales) y —relevante para efectos de este artículo— en el derecho penal (Araya, 2020: 262-266). En efecto, esta última vinculación es una realidad que se ha descrito como tangible desde hace décadas (Hernández, 2019: 794).

5. Sobre la sujeción de la IA al derecho penal, De la Cuesta plantea que «si los entes artificiales inteligentes *actúan* o *realizan comportamientos* con significado social, su *comportamiento y actuación* habrán de adecuarse a normas» (2019: 55).

De esta manera, en una primera dimensión, la IA tiene un impacto potencial en el sistema judicial punitivo, especialmente en materia de prevención e investigación policial de la delincuencia y en la aplicación al proceso de determinación judicial de la responsabilidad penal (Miró, 2018: 97). En este ámbito, se han utilizado los sistemas jurídicos expertos, diseñados «para apoyar la toma de decisiones de los jueces y emitir sentencias en los diferentes juicios que realizan, a partir de un prototipo de sentencia cuya base de conocimiento está integrada por los requisitos de forma y fondo de una determinada sentencia del derecho» (Sánchez y Toro-Valencia, 2021: 215).[6]

Sin embargo, la vinculación de la IA con la rama punitiva no se agota en una funcionalidad puramente auxiliar, en tanto el derecho penal «no puede quedarse al margen de las lesiones a los bienes jurídicos afectados por el uso de esta nueva tecnología» (Valls, 2022: 4). Al respecto, Blanco (2019: 65 y 66) sostiene que hoy en día los robots inteligentes no son los destinatarios de las normas penales, «por lo que, si su decisión ha sido autónoma, sin intervención humana, y en el ejercicio de su propia voluntad, no va a ser objeto de sanción alguna», sin perjuicio que, atendidos los avances tecnológicos, se avizora que ese escenario podría llegar en el corto o mediano plazo.[7] En este contexto, conviene ahondar en cómo responsabilizar penalmente a los agentes humanos que se sirven de la IA (hoy disponible) para la perpetración de ilícitos penales; y solo después aventurarse a escudriñar cómo podría ser la valoración de la responsabilidad jurídico-penal de la *IA fuerte* o *machina sapiens*.

Modelos de atribución de responsabilidad de la agencia humana en relación con la inteligencia artificial

Los robots y sistemas de inteligencia artificial pueden menoscabar bienes jurídicos penalmente protegidos, tales como la salud, la seguridad, el patrimonio, la privacidad o la vida de una persona natural (Amézquita, 2024: 130). Esta afectación puede derivar, por una parte, del empleo deliberado que hace el agente humano al instrumentalizar la IA con fines delictivos; y, por otra, de conductas autónomas —aunque

6. En la experiencia comparada, en 2017 la Fiscalía de la Ciudad Autónoma de Buenas Aires implementó el sistema de inteligencia artificial denominado «Prometea», consistente en un software para preparar dictámenes jurídicos en casos análogos con precedentes judiciales reiterados (Estevez, Linares y Fillottrani, 2020: 10).

7. Con todo, autores como Romeo terminan por aceptar que, llegado cierto punto, «no habría problema en revisar las características actuales de la teoría del delito y adaptarlas para dar cabida a la responsabilidad penal de los sistemas y productos de inteligencia artificial», con la prevención que tal proceder podría «contaminar la teoría del delito aplicable a seres humanos», con el probable efecto de relajar sus requisitos o prescindir de alguno de ellos en detrimento de los *homo sapiens*; problemas que profundizaremos más adelante (Romeo, 2023: 9).

limitadas— ejecutadas por el propio sistema, cuya operación guarda relación funcional con dicho agente (Romeo, 2022: 7 y 8; Valls, 2022: 4).

En lo que respecta a la imputación penal derivada de la agencia humana vinculada a la IA, Valls (2022: 24) distingue tres categorías de intervinientes, conforme a su función dentro del ciclo de vida del sistema: i) diseñadores, fabricantes o programadores del «producto», responsables de su concepción e introducción en el mercado; ii) profesionales que utilizan los sistemas inteligentes (con las particularidades de su interacción, diversa a la de los diseñadores); y iii) los usuarios finales o consumidores.

En este marco —el análisis de la responsabilidad penal de los sujetos humanos involucrados en hechos típicos vinculados a la IA—, es posible categorizar la responsabilidad en los siguientes términos: i) autoría directa del agente humano que utiliza el instrumento delictivo,[8] tanto usuario como diseñador, fabricante o programador; ii) intervención del diseñador, fabricante o programador en hechos ejecutados por terceros; y iii) responsabilidad penal por el desarrollo o circulación imprudente de IA.

Autoría directa mediante el uso de sistemas de inteligencia artificial

Supuestos de utilización directa por el agente humano

En esta circunstancia, la IA puede ser utilizada por un agente como un instrumento para cometer delitos (Muñoz, 2022: 7; Palma, 2023: 252), donde aquella funge como *medio comisivo*[9] a través del cual se realiza o perpetra una conducta tipificada penalmente, tal como sería, en principio, que un agente recurra al engaño, la violencia, la coacción o la intimidación.[10]

En este sentido, el uso de la IA como instrumento delictivo se encuadra en las categorías tradicionales de imputación a título de autoría, donde la conducta del hu-

8. Navarro y Vidal, a modo ejemplificativo, señalan «matar a personas, robar un banco o una defraudación en el mercado de capitales» (2021: 267). Xavier identifica cuatro estructuras delictivas asociadas a la IA: i) delitos cometidos intencionalmente por personas físicas o jurídicas, con uso deliberado de la IA; ii) delitos negligentes causados por fallas en la cadena productiva y/o uso de la IA; iii) ilícitos provocados por la propia IA, sin intervención humana; y iv) ilícitos cometidos por seres humanos, instrumentalizados por la IA (2023: 9).

9. En sede penal, puede distinguirse categorialmente entre los denominados *delitos de medios determinados*, en los cuales la descripción legal contempla expresamente los modos comisivos para la realización del comportamiento incriminado, y los *delitos resultativos*, donde el menoscabo del bien jurídico puede realizarse por cualquier conducta que cause el resultado típico, sin importar el medio empleado (Mir Puig, 2015: 233 y 234).

10. En este sentido, Navarro y Vidal sostienen que la programación de un software de IA para la comisión de fraudes económicos no difiere (sustancialmente) del uso que hace el sujeto activo de un destornillador para forzar la ventana de una casa para sustraer cosas ajenas desde su interior o un cuchillo para fines homicidas (2021: 268 y 269).

mano es replicable a aquellos comportamientos delictivos actualmente sancionados, con la particularidad de que el medio comisivo es la IA (Momblanc, 2024: 197). Así, el sujeto que se vale de un sistema inteligente desplegando una conducta que exhibe las propiedades típicas de un delito, deberá responder como autor —directo o inmediato— de las consecuencias del acto o la causación del resultado (Romeo, 2022: 10; Muñoz, 2022: 9). En estos casos, la IA no actúa de forma autónoma, sino como un vehículo de la conducta del agente humano.

Un ejemplo paradigmático de lo expuesto son los robots utilizados como ciberarmas en contextos de conflictos armados para atacar a la población (Blanco, 2019: 64). Estos, por regla general, son esencialmente teleoperados de manera remota por agentes humanos, por lo cual son considerados como armas avanzadas con escaso poder autónomo (Hellström, 2012: 104). En esta constelación casuística, la atribución de responsabilidad por los resultados dañosos corresponderá directamente al agente humano que utilice o se sirva de los dispositivos IA, sin que se configure un espectro normativo —en términos laxos— de imputación penal a las máquinas (Hellström, 2012: 104; Romeo, 2022: 10).

Supuestos de utilización dolosa de la inteligencia artificial por el diseñador, fabricante o programador

Según lo visto hasta acá, los sistemas de IA presentan una potencialidad para menoscabar bienes jurídicos, ya sea personales o patrimoniales (Blanco, 2019: 63; Xavier, 2023: 8). En este sentido, se ha señalado que, si una IA fuera diseñada o programada con fines delictivos, atendida su capacidad de realizar conductas lesivas, «probablemente los cometería sin error alguno» (Morán, 2021: 303).

Lo expuesto nos conduce a examinar bajo qué condiciones podría atribuirse responsabilidad penal al ser humano que diseña, fabrica o programa a la IA, cuando se verifican resultados típicos (des)valorados penalmente. La primera casuística se compone de aquellas situaciones donde el diseñador, fabricante o programador configura deliberadamente la IA para causar daño o cometer delitos, como el caso de un programador que diseña un software de un robot para que destruya la vivienda de su vecino con quien mantiene una rencilla (Kumar y Kumar, 2019: 18). En este supuesto, el agente utiliza el sistema inteligente que desarrolló para cometer un delito (Araya, 2020: 268 y 269), por lo que se constituye como el perpetrador penalmente responsable.

Para efectos prácticos, se pueden mencionar otros ejemplos: i) un sujeto desarrolla un *malware* con IA para perpetrar fraudes bancarios, lo que constituye eventualmente un fraude informático (artículo 7 de la Ley 21.459); ii) un agente crea con IA una *DeepFake* que genera videos pornográficos falsos, montando el rostro de personas reales, con la finalidad de chantajearlas o atentar contra el honor de las mismas

(Miró, 2025: 187), supuesto que podría eventualmente configurar un delito contra la integridad moral de menores (artículo 403 ter del Código Penal) u otros ilícitos como el de producción y distribución de pornografía infantil (artículo 367 quáter del Código Penal); iii) un diseñador crea un *chatbot* que simula ser un adolescente en redes sociales, para contactar menores de edad, ganar su confianza y obtener imágenes sexuales, hecho eventualmente constitutivo del delito de determinación de menores a la entrega o exhibición de imágenes de significación sexual (artículo 366 quáter del Código Penal); iv) un programador diseña una herramienta de IA entrenada para generar y difundir de forma automatizada y masiva información falsa sobre determinadas sociedades anónimas con el objetivo de distorsionar el precio de las acciones —mecanismo conocido como *scalping* (Londoño, 2023: 97-98)—, hecho que podría ser constitutivo del delito de manipulación informativa del mercado, previsto en el artículo 59 letra f) de la Ley 18.045, de Mercado de Valores.

La responsabilidad del diseñador, fabricante o programador bajo otros títulos de intervención delictiva (coautoría, autoría mediata, inducción y complicidad)

En determinadas hipótesis, donde el diseñador, fabricante o programador de un sistema de IA no ejecuta directamente la conducta típica, su intervención de todas formas podría ser penalmente relevante en función de su vinculación con el hecho cometido por un tercero que emplea la IA para fines ilícitos, en una modalidad de ejecución que no se concreta directamente, sino a través de una instancia donde concurren una pluralidad de agentes (autoría mediata, coautoría, inducción y complicidad).

La *autoría mediata* corresponde a una forma de autoría donde el individuo no ejecuta directamente la conducta que realiza el tipo penal, sino que el quebrantamiento de la norma resulta mediado por el comportamiento de un agente inmediato (el denominado *hombre de adelante*), cuya responsabilidad jurídico-penal es normativamente deficitaria (Kindhäuser, 2011: 48; Mañalich, 2010: 386). Conforme a lo expuesto, se podría configurar esta hipótesis de intervención no ejecutiva en la medida en que el diseñador, fabricante o programador se sirve de un individuo que carece de autonomía decisoria, como sería el caso de un ingeniero que diseña una interfaz de IA para activar remotamente un dron armado y convence a un técnico subordinado de que ejecute una operación de testeo inofensiva, ocultándole que el comando preestablecido en realidad activará un ataque letal contra un blanco humano.

En la *coautoría*, los autores actúan en base a un plan conjunto, donde el comportamiento propio y el ajeno se enmarcan en un esquema común de interpretación (Kindhäuser, 2011: 50 y 51), cuyo actuar implica un aporte funcional al desarrollo de un plan delictivo global, intersubjetivamente representado con otro, que también es (co)autor para estos efectos. Por ejemplo, el caso de un agente que crea un sistema de IA para vulnerar bases de datos médicos, que actúa coordinadamente con un sujeto encargado

de extraer y comercializar la información obtenida, hechos eventualmente constitutivos de delitos informáticos (acceso ilícito previsto en el artículo 2 de la Ley 21.459).[11]

Por su parte, la *inducción* es una forma de participación donde «el hombre de atrás ha motivado determinante el comportamiento del hombre de adelante, pero sin ser competente por un déficit de un presupuesto constitutivo para el delito correspondiente» (Kindhäuser, 2011: 51 y 52). Acá, el diseñador, fabricante o programador de un sistema IA podría ser calificado como inductor si, no estando en condiciones de ejecutar por sí mismo el hecho, motiva o determina al ejecutor para que utilice el sistema por él creado en la perpetración de un delito, influyendo en el proceso deliberativo de un tercero mediante una estrategia persuasiva (González, 2024: 104 y 105). Esto se materializa, por ejemplo, en la decisión de utilizar un sistema de *DeepFakes* para difundir material difamatorio, poniendo a disposición del ejecutor el instrumento delictivo: instruyéndolo directamente sobre cómo utilizarlo, interiorizándolo en la activación de las funciones específicas y garantizando futuros réditos a partir de su empleo.

En el caso de la *complicidad*, el comportamiento del diseñador, fabricante o programador se vuelve causalmente relevante para efectos de la configuración del delito, en virtud de un ulterior comportamiento de un tercero (autor principal), donde la contribución no se enmarca en un plan conjunto representado intersubjetivamente (Kindhäuser, 2011: 48 y 51). En este sentido, resulta necesario que las acciones de apoyo del agente-cómplice aumenten las posibilidades de realización típica para el autor principal. Así, por ejemplo, será cómplice el diseñador, fabricante o programador que entrega un sistema de IA entrenado para vulnerar claves de autenticación, facilitando su uso a un tercero que lo emplea para cometer un fraude informático, aun cuando no exista entre ambos un plan delictivo común, intersubjetivamente representado.

La responsabilidad del diseñador, programador o fabricante de la inteligencia artificial por imprudencia, negligencia o culpa

Miró sostiene que en el marco de la atribución de responsabilidad penal por IA, los casos más problemáticos serán aquellos donde el título de imputación corresponda a la imprudencia (2025: 196). En estos supuestos, no solo se considera la infracción de deberes *ex ante* por parte del desarrollador, sino también los problemas estructurales de atribución causal ante resultados imprevisibles originados por sistemas de inteligencia autónomos. Ambos aspectos se analizan de manera integrada en la medida en que inciden sobre la posibilidad de reproche subjetivo.

11. En tal sentido, se ha planteado que «el método utilizado para la superación de la barrera o medida puede ser de cualquier naturaleza, sea usando programas informáticos, como troyanos o virus, por el método de intento y error o por el uso de preguntas de listas de palabras» (Medina, 2024: 60).

En estos casos, los agentes humanos desarrollan o introducen en el mercado sistemas de IA defectuosos o peligrosos, sin intención delictiva, pero infringiendo deberes de cuidado normativizados. En esta constelación, la irrogación de un daño por la IA se deriva de una actuación negligente o por culpa del agente humano, ya sea por fallas en el diseño, defectos en la programación o errores en la integración de componentes. Por tanto, frente a eventos de esta naturaleza, el análisis deberá enfocarse en «si las fallas del sistema o en su operación obedece a la infracción del deber objetivo de cuidado, ya sea por su creador (defecto de construcción) o el usuario» (Momblanc, 2024: 198). Un ejemplo de lo expuesto es el caso de un vehículo autopilotado que atropella a un peatón menor de edad por una malinterpretación de sus sensores de exploración del entorno, sin identificarlo como un ser humano (Gless, Silverman y Weigend, 2016: 425). En tal hipótesis, la responsabilidad penal recae en el programador, no por dolo, sino por la omisión del deber de cuidado exigible en la configuración del sistema (Kumar y Kumar, 2019: 19).

En dichos casos, el examen de atribución de responsabilidad del diseñador, fabricante o programador respecto de menoscabos típicos de bienes jurídicos puede fundarse en los criterios de la responsabilidad penal por el producto defectuoso (Gless, Silverman y Weigend, 2016: 425; Blanco, 2019: 69), en virtud de los cuales se sancionan las «afectaciones a la vida y salud de los consumidores por la fabricación y puesta en el mercado de productos defectuosos» (Contreras, 2015: 267). Esto, considerando que, por regla general, los sistemas de IA se encontrarán integrados en productos y herramientas de servicio (Gómez, 2025: 9-11).

En el marco del régimen de responsabilidad penal por el producto defectuoso, el diseñador, fabricante o programador puede afectar los bienes jurídicos de los usuarios i) infringiendo deberes de diseño, fabricación o instrucción, o bien, ii) permaneciendo inactivo tras descubrir riesgos del producto no reconocibles al momento de su comercialización (Contreras, 2015: 290). El juicio de reproche o desaprobación jurídica al diseñador, fabricador o programador requiere confrontar la conducta con los estándares normativos exigibles, esto es, con lo mandatado o prohibido por el derecho (Contreras, 2015: 270), examinando si se omitieron o infringieron las medidas necesarias de vigilancia, prevención o retiro del sistema una vez detectado el riesgo.

En este marco, puede extrapolarse el principio que rige en dicha sede de responsabilidad, según el cual cualquier empleo irracional, anómalo o desviado del natural destino de un producto —en este caso, de un sistema de IA— por parte del usuario final, constituye un riesgo no permitido o tolerado por el ordenamiento jurídico (Contreras, 2015: 275). Este criterio excluye la imputación penal al diseñador, fabricante o programador respecto del resultado ulteriormente verificado, siempre que haya actuado sin un propósito delictivo primigenio ni se haya infraccionado deberes posteriores a la puesta en circulación del producto en el mercado. En tales supuestos,

no se verifica una conducta punible ni directa ni mediata por parte de dicho agente (Muñoz, 2022: 10).

En la misma dirección se pronuncia Araya, quien, considerando ciertas características de los sistemas de IA —como su autonomía, imprevisibilidad y la limitada capacidad de control humano—, sostiene que, en principio, no cabría atribuir responsabilidad penal al desarrollador, en tanto «no desplegó ninguna conducta punible» (2020: 268). Esta exclusión se ve reforzada por las dificultades estructurales de control que enfrentan quienes intervienen en la cadena de diseño y fabricación, lo que obligará a mirar responsabilidades en las distintas fases del ciclo de vida de la IA.

A ello se suma el factor del grado de autonomía funcional del sistema. En principio, el nivel de autonomía y capacidad de una IA estará definido y delimitado por el agente humano diseñador, fabricante o programador (Momblanc, 2024: 201). Ahora bien, cuando la IA incorpore capacidades de aprendizaje automático que le permitan modificar su comportamiento a partir de experiencias, datos o interacciones posteriores a su diseño, podrán generarse resultados típicos imprevisibles, incluso para sus propios creadores (Pérez-Arias, 2023: 175; Araya, 2020: 261; Xavier, 2023: 8), dada la complejidad de un comportamiento emergente de la IA (Dremliuga y Prisekina, 2020: 257 y 260), donde el agente humano no está en condiciones de comprender o explicar íntegramente los procesos que conducen a determinadas acciones (Valls, 2022: 23). Si, además, el agente humano no conserva capacidad real de intervención sobre ese comportamiento posterior, la atribución de responsabilidad penal queda excluida (Romeo, 2022: 11), en tanto los resultados típicos derivarán de conductas autónomas e imprevisibles del sistema.[12] En efecto, el riesgo que se concrete en el resultado no sería cognoscible ni evitable conforme a estándares jurídicos de previsibilidad (Miró, 2025: 195; Momblanc, 2024: 198).

Dicho esto, en el estado actual de la ciencia y del desarrollo tecnológico de la IA, no es posible sostener que estos sistemas cuenten con una autonomía estructural completa que los desvincule enteramente del agente humano que los diseña, fabrica o programa (Navarro y Vidal, 2021: 265), por lo cual se sostiene que la imputación penal por hechos derivados del actuar de una IA debería reconducirse —en principio— al agente humano responsable de su diseño, fabricación o programación (Blanco, 2019: 66),[13] descartándose una responsabilidad autónoma del sistema.

12. En este sentido, en el marco de la ciberdelincuencia, véase Pérez-Arias (2023: 180). Por su parte, Gless, Silverman y Weigend sostienen que el hecho de que los agentes inteligentes sean generalmente imprevisibles no puede eximir de responsabilidad a sus operadores, porque es la propia imprevisibilidad de los robots la que da lugar a los deberes de diligencia (2016: 427).

13. Por su parte, Navarro y Vidal plantean que «los conflictos de naturaleza penal que tal tecnología pueda provocar son resueltos aún por medio de la imputación de los resultados dañosos a tales creadores o usuarios» (2021: 265).

Valoración del uso de la inteligencia artificial como medio comisivo y circunstancia agravante de responsabilidad penal

En términos legislativos, el Código Penal chileno no contiene disposiciones que expresamente contemplen o refieran a la IA en algún tipo penal; o bien, que consideren su uso como una circunstancia agravante de responsabilidad. Con todo, el debate normativo ha comenzado a proliferar. Algunos legisladores, advirtiendo el avance de las nuevas tecnologías, han propuesto tipificar el uso de inteligencia artificial en la comisión de un delito como circunstancia agravante de responsabilidad, mediante la incorporación de un numeral 23 en el artículo 12 del Código Penal (boletín 16.021-07, de 13 de junio de 2023, y desde el 6 de mayo de 2025, en primer trámite constitucional del Senado). Por otra parte, el boletín 15.935-07, de 15 de mayo de 2023, propone reformar los tipos de fraudes por engaño, contemplando el uso de sistemas de IA entre los medios típicos, equiparando su uso al nombre fingido, atribución de poder, influencia, crédito, negociación imaginaria, modalidades previstas en el delito de estafa del artículo 468 del Código Penal.

Lo expuesto, en consecuencia, da cuenta de dos técnicas legislativas en cuanto a la valoración jurídico-penal de la IA: por una parte, como una circunstancia agravante de responsabilidad penal; por otra, su inclusión expresa como un medio comisivo en los tipos penales (Miró, 2025: 201 y 202). Sobre esta última técnica, es decir, la incorporación de la IA como medio de comisión alternativo, cabe preguntarse si, en el estado actual de nuestro ordenamiento punitivo, el uso de IA en la comisión de delitos puede ya subsumirse en los tipos penales vigentes. La pertinencia de su incorporación expresa dependerá, en parte, de la casuística y de las exigencias de precisión que impone el principio de legalidad.[14]

A modo de ejemplo, algunos delitos económicos reformados por la Ley 21.595, de Delitos Económicos, como el artículo 468 del Código Penal, presentan redacciones más amplias, lo que allana el camino para identificar —según cada caso— el uso de la IA como circunstancia fáctica ejemplificativa de la propiedad del engaño típico en la estafa. Véase el caso que se produjo en Hungría el 2019, cuando una persona, valiéndose de la IA, realizó una simulación de la voz de un director ejecutivo para solicitar un depósito de 243 mil euros desde una empresa eléctrica a un proveedor, el cual efectivamente se produjo (Morán, 2021: 299). Aquí, la IA funge como un instrumento defraudatorio, esto es, un «medio del que se vale el autor del delito para instrumentalizar al disponente y provocar el perjuicio patrimonial de este o de un

14. El mandato de determinación exige la descripción con mayor certeza posible de la conducta prohibida o mandatada, con el fin de proteger al ciudadano de la arbitrariedad (Kindhäuser y Zimmerman, 2024: 71).

tercero» (Mayer, 2014: 1025). Esto evidencia que puede valorarse a la IA como un medio determinado subsumible en un tipo penal vigente.

En el sentido apuntado, encontramos otros ejemplos en delitos de estructura resultativa, donde no existe una restricción del medio comisivo a emplear. Véase el caso de la utilización de un vehículo aéreo no tripulado que atente contra personas, decantando en un resultado de muerte (Quintero, 2017: 5). De allí que la expresa incorporación de la IA como un medio comisivo tampoco se torna necesaria en este tipo de delitos. Un caso distinto es el del delito de comercialización y producción de material pornográfico infantil, previsto y sancionado en nuestro ordenamiento en el artículo 367 quáter del Código Penal. En virtud de su redacción, resulta discutible si el tenor del tipo penal podría no comprender suficientemente la utilización de la IA para efectos delictivos,[15] por lo que resultaría sugerible su inclusión expresa como modo comisivo.[16]

Ahora bien, un aspecto diverso es la valoración de la IA como una circunstancia agravante de responsabilidad penal, esto es, como un factor de alteración de la pena señalada en el delito (Matus y Ramírez, 2021: 601), tal como lo propone el boletín 16.021-07 («cometer el delito mediante el uso o por medio de inteligencia artificial»).[17] En este sentido, la IA presenta ciertas particularidades que permiten fundar una (des) valoración penológica aumentando su reprochabilidad (Cury, 2011: 498): i) la sofisticación del medio comisivo (como el *DeepFake*), que aumenta su eficacia; ii) la reducción de la capacidad defensiva de la víctima, lo que podría asimilarse con una agravante por mayor facilidad de comisión, como la alevosía; iii) la difuminación de la identificación del agente humano, lo que tiene efectos en la persecución penal; y iv) el efecto multiplicador o intensificador del daño, ya que puede reproducir resultados a gran escala, pudiendo aumentar los resultados lesivos.[18]

15. Aunque la subsunción típica no sea prístina, podría plantearse que la generación de videos pornográficos en los que el agente, valiéndose de IA, haya simulado actividades sexuales de menores, utilizando por ejemplo la voz o imagen de estos (por ejemplo, con un *deepfake*), sería punible según lo dispuesto en el inciso final del artículo 367 quáter del Código Penal. Con todo, cualquiera sea la edad del ofendido, este siempre podría recurrir al tipo penal de injurias, entendido —*lato sensu*— como un delito de aplicación «general o subsidiaria» al interior del catálogo de delitos contras las personas (Mañalich, 2020: 3-17).

16. Recientemente (19 de mayo de 2025), en Estados Unidos fue aprobada la ley «Take it Down Act», que precisamente criminaliza la publicación no consentida de representaciones visuales íntimas, tanto auténticas como generadas artificialmente. Para más información, véase https://tipg.link/gQ3h.

17. El Código Penal peruano, por ejemplo, fue reformado por la Ley 32.314, de 29 de abril de 2025, la cual introdujo a la IA como agravante, pero también en tipos penales específicos, como el previsto en el artículo 129-M, de pornografía infantil, que sanciona la producción de material pornográfico infantil cuando el contenido es generado por inteligencia artificial.

18. En este último sentido, véase Miró (2025: 201).

Lo expuesto, en consecuencia, podría fundar razonablemente la configuración de la IA como una circunstancia modificatoria de responsabilidad agravante, de carácter común —efecto regido por las disposiciones generales de determinación de pena de los artículos 65 a 68 del Código Penal— y material —referida al medio utilizado para ejecutar el hecho (Rodríguez, 2011: 411).

Sin perjuicio de las reflexiones precitadas, el aspecto nuclear es que la IA es una tecnología en expansión, con incidencia en la responsabilidad penal —ya sea como medio comisivo específico o agravante/calificante— y que debe ser tomada en cuenta por el legislador, particularmente para fortalecer el principio de legalidad en su vertiente de mandato de determinación. Las consideraciones hasta acá expuestas, sin embargo, difieren sustancialmente del problema relativo a la eventual responsabilidad penal *autónoma* de la IA, cuestión que se aborda a continuación.

Estado del arte respecto del debate comparado sobre la plausibilidad de castigar penalmente a la *machina sapiens criminalis* o inteligencia artificial fuerte

Efectuado el análisis de los modelos de atribución de responsabilidad penal a la agencia humana, subsiste un amplio debate a nivel comparado sobre la posibilidad de que determinados sistemas de IA puedan ser considerados penalmente responsables por sí mismos. Un sector doctrinario rechaza tal propuesta. En este sentido, sobre la posibilidad de castigar una IA fuerte, del Rosal Blasco sentencia que «la personificación de los sistemas de IA es, a efectos penales, perfectamente inútil porque estos carecen de capacidad de acción, culpabilidad y pena en términos penales» (2023: 43), lo que es secundado por cierta doctrina latinoamericana (Momblanc, 2024: 203 y 204).

En la misma orientación, Osmani funda su negativa en la noción de «autoconciencia», cuyo fundamento se encontraría en la habilidad de las personas (humanas) de pensar y tomar decisiones morales, como lo sería discernir entre el bien y el mal, a partir de lo cual —en su concepto—, desde una perspectiva ética y legal, sería inapropiado hacer penalmente responsable a una IA por sus acciones, ya que no serían conscientes de las consecuencias de sus actos, por lo que tales victimarios —en definitiva— deben considerarse carentes de culpabilidad jurídico-penal (Osmani, 2020: 58).

En similar dirección, Abbott y Sarch consideran que castigar a la IA fuerte en última instancia no estaría justificado, porque traería aparejados costos significativos y requeriría de cambios normativos radicales, siendo mejor enfocarse en ampliar la responsabilidad penal y civil de las personas (naturales) que se sirven de las IA para cometer delitos (2019: 323); salvo que, como sociedad, estemos dispuestos a concederle una «personalidad legal» a la IA, cuestión a la que se oponen tajantemente por una miríada de razones (2019: 375 y 378).

Otra línea argumental se puede visualizar en Lima, quien sin desconocer que ya no es una ficción que una IA potencialmente se vuelva autónoma e independiente de los humanos, de todas maneras rechaza (tomando como base la tecnología disponible en la actualidad) utilizar el derecho penal para su castigo (2018: 695 y 696). Con todo, manifiesta que a futuro podría aceptarse su utilización, en la medida en que la IA (fuerte) sea suficientemente parecida a los humanos, o, alternativamente, instando a revisitar los conceptos de *mens rea* («culpabilidad») y reproche para entidades «no humanas», ya que como mínimo habría una discusión abierta en este punto, por cuanto el concepto de culpabilidad «refleja nuestra experiencia colectiva de lo que significa ser humano» (695 y 696).

También otros autores muestran una flexibilidad semejante. Así, frente a la pregunta de si puede una entidad artificial inteligente ser penalmente responsable de un delito, Blanco sostiene que «la respuesta hoy en día es no», salvo que «se le reconozca personalidad jurídica (electrónica) al robot inteligente» (2019: 73). Por su parte, Gless, Silverman y Weigend defienden que, tal como conocemos a los agentes inteligentes hoy en día, no pueden considerarse «personalmente» responsables por los daños que puedan causar, por ser entidades sin conciencia ni reflexión sobre la naturaleza de sus acciones. No obstante, reconocen que es un asunto posible de replantear, si en un futuro la IA adquiere capacidad de autorreflexión y un equivalente de conciencia (2016: 416 y 417). En el polo opuesto, autores como Kirpichnikov y otros suscriben a la idea de la responsabilidad autónoma de la IA fuerte, al considerar que esta es «distinguible por tener la habilidad de autoanalizarse y por su comportamiento volitivo consciente, por esta razón, esta clase de IA es considerada como un sujeto de responsabilidad penal», sin entrar mayormente en el debate ético planteado por sus contendores (2020: 4).

Según todo lo expuesto hasta acá, es indubitado que una entidad que se deje identificar como una IA fuerte o *machina sapiens criminalis* no es —todavía— una realidad. Pero su eventual llegada, por más remota que parezca, nos invita a reflexionar sobre la necesidad de castigarla autónomamente, haciéndonos cargo de las prevenciones anotadas.

La responsabilidad penal de la persona jurídica como base dogmática para imputar a la *machina sapiens*

La atribución de responsabilidad penal a entes no humanos —como eventualmente podría ser el caso de una IA fuerte— constituye hoy una realidad consolidada en diversos ordenamientos. En la mayoría de los países de referencia (España y Estados Unidos, entre otros) las personas jurídicas, que carecen de voluntad y acción propias, pueden ser sancionadas penalmente con independencia de las personas físicas que las dirigen, bajo la consideración de que representan «fuentes de peligro para bie-

nes jurídicos de terceros» (Hernández, 2024: 46-50). Este reconocimiento relativiza las objeciones estrictas que niegan de plano la posibilidad de que un ente artificial autónomo (aún por llegar) responda penalmente por sus actos, analogía plausible dado que la discusión sobre la responsabilidad de la persona jurídica ya ha quedado superada.

Si bien desde un punto de vista ontológico una IA fuerte podría asemejarse más a un ser humano que a una persona jurídica, lo cierto es que la experiencia dogmática acumulada en torno a la responsabilidad penal corporativa permite identificar criterios útiles aplicables, *mutatis mutandis*, a entes artificiales. La analogía, por tanto, no se funda en una equivalencia esencialista, sino en la funcionalidad y viabilidad de una atribución de responsabilidad penal a entes no humanos.

En esta línea, cabe advertir que los criterios elaborados para la imputación penal individual no deben aplicarse de forma automática a entes ficticios (Artaza, 2024: 103). De ello se sigue que el análisis sobre la responsabilidad penal de la IA fuerte no puede quedar condicionado de manera estricta por las exigencias diseñadas para el sujeto humano. Incluso si esta se asemeja funcionalmente a un ser humano, ello no distorsionaría de manera irremediable el antropomorfismo estructural del derecho penal (Wilenmann y Schürmann, 2024: 127), dado que los límites entre humanos y máquinas son cada vez más difusos y estas últimas podrán contar con capacidades equiparables (Navarro y Vidal, 2021: 270).

A partir de la jurisprudencia comparada[19] sobre la doctrina organicista —que asimila a la persona jurídica con un cuerpo humano (Wilenmann y Schürmann, 2024: 129)—, resulta plausible sostener que una homologación similar, e incluso más intensa, pueda predicarse de una IA fuerte. Por tanto, el aparato conceptual desarrollado para la responsabilidad penal de las personas jurídicas constituye un punto de partida válido para rechazar, al menos preliminarmente, las posiciones que niegan de plano la discusión sobre la responsabilidad penal de la *machina sapiens*.

La atribución de responsabilidad penal a la *machina sapiens*: Problemas y vías de reconstrucción desde la dogmática penal

Problemas de la teoría del delito tradicional ante una inteligencia artificial fuerte

En el derecho penal continental tradicional, el delito se concibe bajo una estructura tripartita (tipicidad, antijuridicidad y culpabilidad); aunque es de notar que la tendencia comprende que las dos primeras categorías (tipicidad y antijuridicidad) constituyen el denominado injusto penal, en tanto la imputación personal (culpabilidad) analiza en qué condiciones es posible atribuir un hecho típico y antijurídico al autor

19. Véase *Tesco Supermarkets Ltd. v Nattrass* [1971] UKHL 1 (31 de marzo de 1971).

(Rettig, 2019: 27). En adelante, el análisis se enfocará en la tipicidad y culpabilidad (y su vinculación con los hechos ejecutados autónomamente por la IA fuerte), dejando de lado la antijuridicidad. Sintéticamente, el tipo penal está compuesto por elementos objetivos (descriptivos y normativos) y subjetivos que permiten identificar cada delito de la parte especial del derecho penal. La tipicidad, por su parte, consiste en la adecuación de la conducta a las propiedades que constituyen el tipo (Rettig, 2019: 31).

Ahora bien, en lo referido a la satisfacción de los elementos objetivos y subjetivos del tipo penal por un hecho autónomo de la *machina sapiens*, existe doctrina que no vería impedimento en aquello. Así, partiendo de la base de lo dispuesto en el artículo 1 del Código Penal chileno, se ha afirmado que «debe aceptarse que es el propio ordenamiento jurídico el que decide quiénes o qué entidades pueden ser penalmente responsables» y que «el legislador, normativamente, no ha incluido ninguna restricción en el sentido que tales conductas solo pueden ser ejecutadas por seres humanos», razón por la cual, en última instancia, la responsabilidad penal se «definirá en la medida que tales entidades cumplan con los presupuestos, esto es, que sean capaces de decidir voluntariamente la ejecución de una acción o una omisión» (Navarro y Vidal, 2021: 270 y 271).

En esta línea, se ha sostenido que la tipicidad penal, por regla general, prescinde de elementos como la emotividad o la conciencia moral, exigiendo un grado de libertad relativamente débil. Así, mientras se verifiquen actos dotados de conocimiento e intención, el derecho penal no requiere una voluntad «libre» en sentido filosófico, lo que permitiría —al menos en términos abstractos— no excluir *a priori* la adecuación típica de una conducta ejecutada por una IA fuerte (Navarro y Vidal, 2021: 273 y 274).

Respecto de la culpabilidad, dentro del esquema tradicional que entiende el delito como una acción (u omisión) típica, antijurídica y culpable, este último elemento categorial ha presentado una notable evolución con relevantes implicancias sistemáticas. Así, es posible advertir que la culpabilidad, como elemento del ilícito penal, trasunta —según se expone tradicionalmente— por tres teorías fundamentales:[20] i) la culpabilidad psicológica; ii) la culpabilidad normativa (compleja); y iii) la culpabilidad normativa (pura o finalista) (Garrido Montt, 2007: 259), sin perjuicio de las propuestas de los nuevos modelos dogmáticos.

El esquema finalista contemporáneo, sin los elementos del dolo y la culpa, que fueron trasladados a la tipicidad, se entiende como un poder actuar de otro modo (Yáñez, 1994: 1197) y exige la satisfacción de tres condiciones copulativas: i) imputabilidad; ii) consciencia de la antijuridicidad; y iii) exigibilidad de otra conducta[21] (Garrido Montt, 2007: 263; Künsemüller, 2001: 207; Vargas, 2011: 168).

20. A mayor abundamiento sobre la evolución histórica de la culpabilidad, véase Yáñez, 1994: 1177 y ss.; y Náquira, 2015: 28 y ss.

21. El neokantismo ya incluía i) y iii). Véase Etcheberry, 1998: 272.

A modo de referencia, la imputabilidad dice relación con la aptitud o capacidad que tiene el individuo para comprender la trascendencia jurídica de su actuar y de poder autodeterminarse libremente conforme a derecho (Garrido Montt, 2007: 270). Por su parte, la consciencia de la ilicitud (o de la significación de antijuridicidad del hecho) estaría vinculada a la posibilidad de comprender que tiene el sujeto imputable, en la situación concreta, la (i)licitud de su conducta (Garrido Montt, 2007: 270). Por último, la exigibilidad corresponde a la posibilidad, determinada por el ordenamiento jurídico, de obrar en una forma distinta y mejor que aquella por la que el sujeto se decidió (Cury, 2011: 449).

Frente a estas exigencias, surge la duda si una IA fuerte o *machina sapiens* sería capaz de culpabilidad. Para abordar esta interrogante, resulta útil revisar la manera en que se ha enfrentado este presupuesto respecto de otros entes no humanos, como las personas jurídicas.

Culpabilidad de la persona jurídica como base funcional para la inteligencia artificial

En cuanto a la idea dominante de culpabilidad, se ha señalado que pareciera imposible realizar una aplicación en una persona jurídica de los requisitos de imputabilidad, exigibilidad de la conducta y consciencia de la ilicitud, tal como se aplican a la persona natural (Navas, 2018: 1041). Entonces, existirían dos alternativas: o renunciar a la culpabilidad del agente no humano (IA fuerte) en tanto problema exclusivo de la responsabilidad penal individual, o bien, como se plantea en sede de responsabilidad de la persona jurídica, «entender que la culpabilidad, en este ámbito, también cumple el propósito de "reprochar" la realización del injusto, sin perjuicio de que tal elemento pueda construirse en base a criterios diversos al de la responsabilidad penal individual» (Artaza, 2024: 104).

De inclinarse por la segunda opción, fundada en otros criterios *ad hoc* para el agente no humano, entonces podría el elemento *culpabilidad* disociarse del contenido de garantía vinculado a la responsabilidad penal individual (Artaza, 2024: 112). De esta manera, si el derecho penal ha admitido que la persona jurídica —entidad sin consciencia— pueda ser considerada culpable a partir de un juicio normativo sobre su estructura organizativa y capacidad de autorregulación, entonces, *a fortiori*, podría sostenerse que una IA fuerte, dotada de una autonomía decisional, autoaprendizaje y operatividad independiente, también puede ser evaluada desde una lógica similar. Si una IA fuerte actúa en base a un diseño operativo que le permite adoptar decisiones bajo ciertos estándares normativos —y si estos pueden ser calibrados, supervisados o verificados—, entonces cabría preguntarse si la omisión de tales controles o el «modo de ser» operativo de la IA constituyen, en sí mismos, una base de reproche. La comparación con la persona jurídica, en este sentido, permite disociar la culpabilidad de

un contenido psicológico tradicional, reemplazándolo por una valoración estructural de la agencia artificial.

A diferencia de la persona jurídica y la objeción de castigo en su contra, la IA fuerte sí se encontrará dotada de la capacidad de adoptar posiciones frente a una norma (probablemente, en mejores condiciones que cualquier humano). Debemos considerar que el examen —hipotético— se está efectuando en relación con una entidad que asimilará y procesará información de manera autónoma, con una capacidad decisoria para orientar su comportamiento. De allí que la IA fuerte incluso podrá afrontar un escenario más adecuado para someterla a un examen de culpabilidad: por ejemplo, a través de un *input* podrá ser asequible para ella la totalidad del derecho, en tanto conjunto de normas permisivas y prohibitivas en virtud de las cuales debe adecuar su comportamiento, lo que excluiría la aplicabilidad de un error de prohibición a su respecto (Cury, 2011: 439). A su vez, por tratarse de una IA fuerte, es posible que sus condiciones de programación, aprendizaje y contexto le otorguen una posibilidad técnico-funcional de motivarse conforme a la norma, internalizando patrones de decisión vinculantes. De ser ese el caso, en términos valorativos, si consideramos comunicativamente como destinataria de las normas a una persona jurídica, también podríamos hacerlo con la IA fuerte.

Adicionalmente, siempre y cuando el legislador democrático esté dispuesto a reconocer una «personalidad legal/artificial», resulta atingente la siguiente justificación: «A quien se le concede personalidad debe asumir, como contracara, las cargas que ello implica, incluyendo la posibilidad de hacerse responsable y aplicarle las sanciones que correspondan» (Wilenmann y Schürmann, 2024: 123).

Por otra parte, se ha sostenido que «la culpabilidad de la persona jurídica debe ser entendida como una culpabilidad por el carácter» (Mañalich, 2011: 302), entendiendo por «carácter» la aptitud de exhibir un determinado sello, esto es, una configuración funcional relativamente estable que expresa tendencias o disposiciones a actuar de cierto modo. A diferencia de las personas naturales, a las personas jurídicas no se les reconoce personalidad moral, por lo que el derecho puede censurar directamente su configuración funcional como tal, incluso mediante su supresión institucional, sin incurrir en contradicción (Mañalich, 2011: 302).

Trasladando ese razonamiento al caso de la *machina sapiens*, podría sostenerse, al menos en términos provisionales, que su eventual carencia de dignidad intrínseca —cuestión debatible si es que la IA fuerte constituirá «una nueva especie en la tierra», la que, salvo sus componentes artificiales, en poco más se diferenciaría de un humano— no impediría construir un reproche penal dirigido a su carácter operativo, entendido como un conjunto de disposiciones conductuales estables. En la medida que esa configuración se revele persistentemente disfuncional o antisocial, consideramos que no existiría obstáculo dogmático para imputarle responsabilidad penal bajo un modelo análogo al desarrollado para las personas jurídicas. Así, si la ontología fun-

cional de la *machina sapiens* expresa un modo de ser objetivamente reprochable, el derecho podría válidamente sancionarla a través del expediente penal, del mismo modo en que hoy se admite respecto de las entidades corporativas. De ahí que se pueda concordar con quienes sostienen que «la línea divisoria entre una persona jurídica y una IA es muy delgada» (Chandra y Sanjaya, 2023: 61).

De esta manera, la comparación con la persona jurídica permite disociar la culpabilidad en un sentido psicológico tradicional, reemplazándola por una valoración estructural de la agencia artificial. Así, compartimos en términos generales la reflexión a la que arriba parte de la doctrina comparada, que plantea que «si todos los requisitos específicos de la responsabilidad penal aplicables a los seres humanos pueden extenderse a las empresas, no hay razón por la que no puedan aplicarse también a las entidades de IA» (Kumar y Kumar, 2019: 20).

Ahora bien, la propuesta funcional descrita será reforzada en la sección siguiente al revisar el modelo de imputación elaborado por Hallevy, aplicable a entidades artificiales bajo una lógica penal anglosajona.

El modelo de Hallevy desde el *common law*: Hacia una culpabilidad artificial

En términos generales, la jurisprudencia anglosajona estructura los delitos de forma bipartita, distinguiendo entre el *actus reus* («la porción externa o física del delito») y el *mens rea* («el rasgo mental o interno») (Dressler, 2022: 83). Como ha resuelto la Corte Suprema de los Estados Unidos, la imposición de responsabilidad penal requiere pruebas de «una mente malintencionada, y de manos malhechoras».[22]

Ahora bien, respecto del primero de estos elementos, para la doctrina mayoritaria de dicho país, el *actus reus* se forma de tres componentes: i) un acto voluntario (o, raramente, fallar en actuar), ii) que causa, iii) un daño social (Dressler, 2022: 83). En un sentido semejante, se ha definido al *actus reus* como «la ejecución de algún acto voluntario prohibido por la ley» (Kadish y otros autores, 2022: 477). Sintéticamente, se ha entendido por acto «cualquier movimiento corporal ejecutado voluntaria o involuntariamente» y, por su parte, la voluntariedad de un acto «involucra el uso de una mente humana» (Dressler, 2022: 86 y 88).

Por su parte, el *mens rea* permite ser entendido en dos formas: i) en sentido amplio, como el fundamento moral para declarar culpables a las personas y castigarlas (Bergman y Bergman, 2020: 247), o, análogamente, la reprochabilidad *(blameworthiness)* moral que justifica un castigo (Dressler, 2022: 116); y ii) en sentido estricto, donde se considera que el *mens rea* «es el estado mental culpable del hechor con respecto

22. Véase sentencia *Morissette v. United States*, 342 U.S. at 250, Corte Suprema de los Estados Unidos.

al resultado de su conducta», o «simplemente el particular estado mental descrito en la definición de un delito» (Dressler, 2022: 89 y 117).[23]

Ahora, recogiendo la base conceptual mencionada, de forma pionera, Hallevy traslada el análisis del *actus reus* y el *mens rea* al hecho ejecutado por una *machina sapiens criminalis*, y concluye que no se verifican mayores inconvenientes para atribuirle responsabilidad penal. En efecto, de llegar a existir en el futuro (quizás no tan lejano) esa IA fuerte, no cabría duda de que podría actuar voluntariamente causando un daño social, dado que estaría provista de una mente «funcionalmente equivalente» a la humana (Hallevy, 2013).

Si Hallevy llevase la razón en este punto, esa entidad no humana cumpliría con todos los requisitos del *mens rea*. En términos cognitivos, se puede considerar estas IA como seres con conciencia *(awareness)* al absorber datos del mundo exterior a través de sus «sentidos artificiales» (considérese cámaras en vez de ojos, micrófonos en lugar de oídos, termómetros o transductores para sentir la temperatura o la presión, etcétera), creando, procesando o actualizando conocimientos o imágenes a partir de esos datos en su «mente artificial» (Hallevy, 2013: 50 y ss.). A nivel volitivo, la *machina sapiens criminalis*, al tomar una decisión, efectivamente puede desplegar «distintos niveles de voluntad: intención, indiferencia, y temeridad» (Hallevy, 2013: 58 y ss.).

En la misma línea, otros autores también favorecen la posibilidad de que una IA fuerte obre con una «mente culpable», en la medida en que la mayoría de los sistemas legales descomponen la culpabilidad (y la graduación de la responsabilidad) en elementos cognitivos y volitivos, aspectos predicables en el comportamiento de una IA avanzada (Dremliuga y Prisekina, 2020: 257 y 260). En concreto, conciben a esta como una máquina o software que en ocasiones posee amplias capacidades cognitivas (conocer, comprender y pensar) y voluntad autónoma, configurándose como un «arma» que puede decidir y actuar por sí misma, lo cual constituye un reto que la teoría tradicional del derecho penal no ha podido superar (Dremliuga y Prisekina, 2020: 257 y 260). De este modo, queda al menos provisionalmente establecido que una IA fuerte puede obrar con una *mente malintencionada* y *manos malhechoras*.

Del fin de la pena y el castigo a la *machina sapiens*

Si se admite que la *machina sapiens* podrá actuar con un grado de autonomía y capacidad decisional funcionalmente equivalente a la humana, entonces resulta jurídicamente atendible imputarle responsabilidad y aplicar una sanción penal. Esta posibilidad, sin embargo, para ciertos autores estaría condicionada a que la IA pueda *sentir*

23. En términos semejantes, Kadish y otros autores mencionan que: «El *mens rea* en sentido estricto es un requisito más formal y técnico; se refiere al tipo de conciencia mental o intención que debe acompañar al acto prohibido, según los términos de la ley que define el delito» (2022: 533).

el castigo (Gless, Silverman y Weigend, 2016: 435)[24], es decir, que la consecuencia impuesta esté conectada cognitivamente con el hecho que lo motivó.

En el derecho anglosajón, el castigo (pena) suele asociarse a cuatro posibles fines: retribución, prevención, rehabilitación e incapacitación (Hallevy, 2013: 156 y ss.).[25] En nuestra tradición continental, los fines de la pena se identifican principalmente con la prevención general o especial —sean negativas o positivas— y la retribución.[26]

Para Hallevy, retribución y prevención no serían útiles para castigar la *machina sapiens criminalis*, ya que no experimentan sufrimiento ni miedo (2013: 168). Por ello, solo quedarían la rehabilitación y la incapacitación. No obstante, esta conclusión ha sido cuestionada: si la IA fuerte es autoconsciente, también podría ser capaz de experimentar sufrimiento (Venezian, 2022: 2). En ese caso, una justificación plausible sería la prevención especial negativa, que busca impedir la reiteración del delito mediante el aislamiento de la fuente de riesgo, que en este caso correspondería a los sistemas inteligentes (Romeo, 2022: 13).

Se han propuesto tres etapas sucesivas para establecer penas a la IA fuerte: i) identificar el significado esencial de la pena en cuestión (privación de libertad, muerte, multa); ii) trasladar ese significado al contexto de la IA; y iii) homologar la sanción concreta que se aplicará (Hallevy, 2013: 162 y ss.).

Sobre esa base, se han formulado las siguientes penas aplicables a la *machina sapiens*: i) muerte, como eliminación permanente del software o hardware (existiendo fervientes detractores de la pena de muerte para la IA fuerte, como Lemley y Casey, 2019: 1391); ii) prisión o encarcelamiento *(imprisonment)*, entendida como privación funcional de libertad mediante medidas como el monitoreo permanente y la restricción de actividades, similares a castigos aplicados en Estados Unidos a las corporaciones responsables de quiebras fraudulentas; y iii) las multas pecuniarias, concebidas como «una contribución forzada de patrimonio valioso en favor de la sociedad», lo cual se lograría obligando a que la IA dedique horas de trabajo que generen suficientes ingresos para palear el daño causado (Hallevy, 2013: 165 y ss.).

Similarmente, también se sugieren como sanciones la: i) destrucción física del robot (equivalente a una sentencia de muerte); ii) destrucción o reescritura de algoritmos morales del robot (equivalente a una internación hospitalaria); iii) inmovilización funcional del sistema (privación de libertad); y iv) ordenar multas con cargo al fondo de seguro (Hu, 2019: 529).

---

24. En este sentido, Chandra y Sanjaya afirman categóricamente que la «*inteligencia artificial emocional* ya no es ciencia ficción», pudiendo la IA autónoma sentir dolor, por ejemplo, mediante el borrado de su memoria o la supresión de datos (2023: 60).

25. Sobre las teorías de legitimación de las normas de sanción, véase Kindhäuser y Zimmerman, 2024: 61 y ss.

26. En nuestro medio, la conceptualización más depurada del retribucionismo es aquella ofrecida por Mañalich (2018: 38 y ss.).

Se ha considerado también la posibilidad de aplicar penas sustitutivas (Araya, 2018: 25),[27] tales como la libertad vigilada *(probation)* o la prestación de servicios en favor de la comunidad (cuantificable temporal y económicamente); siempre y cuando cumplan análogamente con los mismos presupuestos que se exigen para los individuos de la especie humana (Hallevy, 2013: 169 y ss. y Jhudele, 2016: 20 y ss.).

Otra alternativa sería desestimar el uso de penas en sentido estricto (debido a los problemas en sede de culpabilidad) y considerar únicamente el uso de medidas de seguridad contra una IA fuerte, esto es, consecuencias jurídicas del delito, distintas de la pena, consistentes en «la privación o restricción de bienes jurídicos, fundada en la peligrosidad criminal del sujeto, con exclusiva función de prevención especial» (Falcone, 2007: 237).

No obstante, esta posición es objetable si se considera que las medidas de seguridad están destinadas a sujetos peligrosos que no pueden responder penalmente (como los inimputables) y que no ofrecen garantías en su comportamiento, por lo que —según se ha propuesto— deben ser combatidos por su contrariedad (desviación) permanente del derecho vigente (terroristas, según Jakobs, 2003: 56). Ese no será siempre el caso de la *machina sapiens*, por lo que las medidas de seguridad solo deben aplicarse bajo condiciones análogas a las vigentes para humanos, conforme a la naturaleza del agente y el hecho cometido.

Conclusiones

De acuerdo con lo desarrollado hasta acá, la inteligencia artificial —en particular aquella denominada *fuerte*, dotada de un alto grado de autonomía funcional, capacidad de aprendizaje y procesamiento— plantea desafíos relevantes para el sistema de imputación penal. El análisis debe reconocer que, bajo las condiciones actuales, la responsabilidad recae sobre los operadores humanos vinculados al diseño, implementación o utilización de sistemas de IA, en la medida en que se satisfagan los presupuestos típicos de los delitos correspondientes, sea a título de dolo o negligencia.

Sin embargo, frente a eventuales escenarios en que la *IA fuerte* opere al margen de todo control humano efectivo, surge la posibilidad de —al menos— considerar una responsabilidad penal autónoma. Para abordar esta hipótesis, se ha recurrido tanto a la estructura del delito desarrollada por la dogmática penal continental como a los aportes provenientes del modelo del *common law*. Asimismo, se han explorado vías conceptuales para fundar el reconocimiento de la culpabilidad, en base al aparataje conceptual dogmático de la responsabilidad penal aplicada a entidades no huma-

27. En el plano local, el autor enfatiza que no puede perderse de vista que las penas sustitutivas son auténticas penas, por cuanto constituyen consecuencias jurídicas negativas, que implican la afectación de derechos fundamentales, impuestas por una sentencia penal condenatoria.

nas, como el de la persona jurídica. Sin perjuicio de que, en términos ontológicos, la *machina sapiens* puede distanciarse del estatuto de la persona jurídica, esta analogía resulta útil como base funcional para explorar vías dogmáticas que sustenten su eventual responsabilidad penal.

En consecuencia, la atribución de responsabilidad penal a sistemas artificiales avanzados no resulta necesariamente incompatible con las estructuras dogmáticas del derecho penal, siempre que los criterios de imputación se redefinan para adecuarse a la naturaleza técnica y operativa de estos agentes no humanos. Ello, por lo demás, no excluye la conservación de un modelo de responsabilidad humana residual o paralela, destinado a evaluar la intervención de quienes diseñan, configuran o utilizan delictivamente estos sistemas. En definitiva, el desarrollo de la *IA fuerte* exige revisar y adaptar los fundamentos normativos de la imputación penal, incorporando esquemas que permitan responder tanto desde la perspectiva del agente humano como de la eventual agencia artificial autónoma.

Referencias

Abbott, Ryan y Alex Sarch (2019). «Punishing artificial intelligence: Legal fiction or science fiction». *UC Davis Law Review*, 53 (1): 323-384. Disponible en https://tipg.link/gx8g.

Amézquita, Jorge (2024). «Inteligencia artificial, riesgos y *compliance*». En Paula Ramírez (directora), *La inteligencia artificial y las nuevas fronteras jurídicas* (pp. 107-138). Valencia: Tirant lo Blanch.

Araya, Luis (2018). *Régimen de penas sustitutivas. Revisión a la Ley 18.216, Ley 20.587 y Decreto Ley 321.* 1.ª ed. Santiago: Der.

Araya, Carlos (2020). «Desafíos legales de la inteligencia artificial en Chile». *Revista Chilena de Derecho y Tecnología*, 9 (2): 257-290. DOI: 10.5354/0719-2584.2020.54489.

Artaza, Osvaldo (2024). *Responsabilidad penal de las personas jurídicas.* 2.ª ed . Santiago: Academia Judicial de Chile y Der.

Barona, Silvia (2019). «Cuarta revolución industrial (4.0.) o ciberindustria en el proceso penal: Revolución digital, inteligencia artificial y el camino hacia la robotización de la justicia». *Revista Jurídica Digital UANDES*, 3 (1): 1-21. DOI: 10.24822/rjduandes.0301.1.

Bergman, Paul y Sara Bergman (2020). *The criminal law handbook. Know your rights, survive the system.* 16.ª ed. California: Nolo.

Blanco, Isidoro (2019). «*Homo Sapiens* y ¿*Machina Sapiens*?: Un derecho penal para los robots dotados de inteligencia artificial». En Covadonga Mallada (directora), *Nuevos retos de la ciberseguridad en un contexto cambiante* (pp. 63-80). Madrid: Thomson Reuters Aranzadi.

Chandra, Rushil y Karun Sanjaya (2023). «Punishing the unpunishable: A liability framework for artificial intelligence systems». En Saad Motahhir y Badre Bossoufi (editores), *Digital technologies and applications, proceedings of ICDTA'23, Fez, Morocco. Tomo 2* (pp. 55-64). Cham: Springer.

Cury, Enrique (2011). *Derecho Penal. Parte general*. 10.ª ed. Santiago: UC.

De la Cuesta, Paz (2019). «Inteligencia artificial y responsabilidad penal». *Revista Penal México*, 9 (16-17): 51-62. Disponible en https://tipg.link/gx8_.

Estevez, Elsa, Sebastián Linares y Pablo Fillottrani (2020). *Prometea: Transformando la administración de justicia con herramientas de inteligencia artificial*. Washington: Banco Interamericano de Desarrollo. DOI: 10.18235/0002378.

Del Rosal, Bernardo (2023). «¿El modelo de la responsabilidad penal de las personas jurídicas para los daños punibles derivados del uso de la inteligencia artificial?». *Revista de Responsabilidad Penal de Personas Jurídicas y Compliance*, 2: 2-49. Disponible en https://tipg.link/gx93.

Dremliuga, Roman y Natalia Prisekina (2020). «The concept of culpability in criminal law and AI systems». *Journal of Politics and Law*, 13 (3): 256-262. DOI: 10.5539/jpl.v13n3p256.

Dressler, Joshua (2022). *Understanding criminal law.* 9.ª ed. Carolina del Norte: Carolina Academic Press.

Fahim, Sadaf y G. S. Bajpai (2020). «AI and criminal liability». *Indian Journal of Artificial Intelligence and Law*, 1 (1): 70-106. Disponible en https://tipg.link/gx9Z.

Falcone, Diego (2007). «Una mirada crítica a la regulación de las medidas de seguridad en Chile». *Revista de Derecho* (Pontificia Universidad Católica de Valparaíso), 29 (2): 235-256. DOI: 10.4067/S0718-68512007000100007.

Garrido, Mario (2007). *Derecho Penal. Parte general*. Tomo 2. 4.ª ed. Santiago: Jurídica de Chile.

Gless, Sabine, Emily Silverman y Thomas Weigend (2016). «If robots cause harm, who is to blame? Self-driving cars and criminal liability». *New Criminal Law Review: An International and Interdisciplinary Journal*, 19 (3): 412-436. DOI: 10.2139/ssrn.2724592.

Gómez, Carlos (2025). «Responsabilidad por daños causados por la inteligencia artificial». *Revista para el Análisis del Derecho,* 3: 9-11. Disponible en https://tipg.link/gWdL.

González, Diego (2024). «Autoría y participación por imprudencia: Posibilidad y límites bajo el Código Penal Chileno». *Revista de Ciencias Sociales* (Universidad de Valparaíso), 84: 83-123. DOI: 10.22370/rcs.2024.84.3837.

Grupo de Expertos en Inteligencia Artificial de la Comisión Europea (2019). «A definition of AI: Main capabilities and disciplines». Disponible en https://tipg.link/gx9x.

Hallevy, Gabriel (2013). *When robots kill: Artificial intelligence under criminal law*. Boston: Northeastern University Press.

Hellström, Thomas (2013). «On the moral responsability of military robots». *Ethics and Information Technology,* 15 (2): 99-107. DOI: 10.1007/s10676-012-9301-2.

Hernández, María (2019). «Inteligencia artificial y derecho penal». *Actualidad Jurídica Iberoamericana*, 10: 792-843. Disponible en https://tipg.link/gxAo.

Hernández, Héctor (2024). «Responsabilidad penal de las personas jurídicas luego de la Ley 21.595. Cambios para la continuidad». En Héctor Hernández, *La Guía del compliance. Responsabilidad penal de las empresas. Modelos de prevención* (pp. 17-85). Santiago: Libromar.

Hildebrandt, Mireille (2020). *Law for computer scientist and other folk*. Oxford: Oxford University Press.

Hu, Ying (2019). «Robot criminals». *University of Michigan Journal Law Reform*, 52 (2): 487-531. DOI: 10.36646/mjlr.52.2.robot.

Jakobs, Gunther (2003). *Derecho penal del enemigo*. 1.ª ed. Madrid: Thomson Civitas.

Jhudele, Priyam (2016). «On robot crimes and punishments». *NLIU Law Review*, 6 (1): 1-25. Disponible en https://tipg.link/gxAE.

Kadish, Sanford, Stephen Schulhofer y Rachel Barkow (2022). *Criminal law and its processes. Cases and materials*. 11.ª ed. Nueva York: Aspen Publishing.

Kindhäuser, Urs (2011). «Infracción de deber y autoría: Una crítica a la teoría del dominio del hecho». *Revista de Estudios de la Justicia* (14): 41-52. DOI: 10.5354/rej.voi14.28552.

Kindhäuser, Urs y Till Zimmerman (2024). *Derecho Penal. Parte general*. 9.ª ed. Valencia: Tirant lo Blanch.

Kirpichnikov, Danila, Albert Pavlyuk, Yulia Grebneva y Hilary Okagbue (2020). «Criminal Liability of the Artificial Intelligence». *E3S Web of Conferences, 159:* 1-10. DOI: 10.1051/e3sconf/202015904025.

Kumar, Ankit y Amit Kumar (2019). «Criminal liability of the artificial intelligence entities». *Nirma University Law Journal,* 8 (2): 15-20.

Künsemüller, Carlos (2001). *Culpabilidad y pena*. Santiago: Jurídica de Chile.

Kurki, Visa (2019). *A theory of legal personhood*. Oxford: Oxford University.

La Parra, Juan Ramón (2021). «Los desafíos de la inteligencia artificial». *Revista Pastoral Juvenil*, 533: 17-30.

Lemley, Mark y Bryan Casey (2019). «Remedies for robots». *University of Chicago Law Review*, 86 (5): 1311-1396.

Lima, Dafni (2018). «Could AI agents be held criminally liable? Artificial intelligence and the challenges for criminal law». *South Carolina Law Review*, 69 (3): 677-694.

Lior, Anat (2020). «AI entities as AI agents: Artificial intelligence liability and the AI respondeat superior analogy». *Mitchell Hamline Law Review*, 46 (5): 1043-1102.

Londoño, Fernando (2023). «Ilícito de manipulación bursátil: Fenómeno y lesividad. Aspectos de política sancionatoria». *Revista Política Criminal,* 8 (15): 64-127.

Mañalich, Juan Pablo (2010). «La estructura de la autoría mediata». *Revista de Derecho* (Pontificia Universidad Católica de Valparaíso), 34: 385-414.

—. (2011). «Organización delictiva. Bases para su elaboración dogmática en el derecho penal chileno». *Revista Chilena de Derecho,* 38 (2): 279-310.

—. (2018). «Retribucionismo consecuencialista como programa de ideología punitiva». En Juan Pablo Mañalich, *Estudios sobre la fundamentación y determinación de la pena* (pp. 25-63). 1.ª ed. Santiago: Thomson Reuters.

—. (2020). «La injuria como delito general contra la persona». En Juan Pablo Mañalich, *Estudios sobre la parte especial del derecho penal chileno* (pp. 3-97). 1.ª ed. Santiago: Thomson Reuters.

Martínez, Goretty (2012). «La inteligencia artificial y su aplicación al campo del derecho». *Alegatos*, 82: 827-846.

Matus, Jean Pierre y Cecilia Ramírez (2021). *Manual de Derecho Penal chileno. Parte general.* 2.ª ed. Valencia: Tirant lo Blanch.

Mayer, Laura (2014). «El engaño concluyente en el delito de estafa». *Revista Chilena de Derecho*, 41 (3): 1017-1048.

McCarl, Ryan (2022). «The limits of law and AI». *University of Cincinnati Law Review*, 90 (3): 923-950.

Medina, Gonzalo (2024). «El delito de acceso ilícito a sistemas informáticos». En Samuel Malamud y Guillermo Chahuán (coordinadores), *Delitos informáticos* (pp. 49-71). Valencia: Tirant lo Blanch.

Mir Puig, Santiago (2015). *Derecho penal. Parte general.* 10.ª ed. Barcelona: Reppertor.

Miró, Fernando (2018). «Inteligencia artificial y justicia penal: Más allá de los resultados lesivos causados por robots». *Revista de Derecho Penal y Criminología*, 3 (20): 87-130.

—. (2025). «Derecho penal y desafíos de la inteligencia artificial (en el contexto del nuevo marco regulatorio europeo». En María Emilia Casas (directora) y Daniel Pérez (coordinador), *Derechos y tecnologías* (pp. 177-216). Madrid: Centro de Estudios Ramón Areces.

Momblanc, Liuver (2024). «Inteligencia artificial y derecho penal. ¿Será necesario un nuevo concepto de delito?». *Lex-Revista de la Facultad de Derecho y Ciencias Políticas*, 34 (22): 185-209.

Morán, Alejandra (2021). «Responsabilidad penal de la inteligencia artificial (IA). ¿La próxima frontera?». *Revista del Instituto de Ciencias Jurídicas de Puebla*, 15 (48): 289-323.

Morillas, David (2023). «Implicaciones de la inteligencia artificial en el ámbito del Derecho Penal». En Jaime Peris y Antonella Massaro (editores), *Derecho penal, inteligencia artificial y neurociencias* (pp. 59-92). Roma: Roma Tre-Press.

Muñoz, José (2022). «Inteligencia artificial y responsabilidad penal». *Derecho Digital e Innovación*, 11: 1-34.

Náquira, Jaime (2015). *Derecho penal. Parte general*. Tomo 1. 2.ª ed. Santiago: Thomson Reuters.

Navarro, Roberto e Iván Vidal (2021). «Sobre la justificación de aplicar el derecho penal a las entidades de inteligencia artificial». En Michelle Azuaje y Pablo Contreras (editores), *Inteligencia artificial y derecho: Desafíos y perspectivas* (pp. 261-281). Valencia: Tirant lo Blanch.

Navas, Iván y Antonia Jaar (2018). «La responsabilidad penal de las personas jurídicas en la jurisprudencia chilena». *Revista Política Criminal,* 13 (26): 1027-1054.

Novoa, Eduardo (2005). *Curso de Derecho Penal chileno. Parte general.* Tomo 1. 3.ª ed. Santiago: Jurídica de Chile.

OCDE, Organización para la Cooperación y el Desarrollo Económico (2022). *Recommendation of the Council on Artificial Intelligence*. Disponible en https://tipg.link/gRKU.

Osmani, Nora (2020). «The complexity of criminal liability of AI systems». *Masaryk University Journal of Law and Technology*, 14: 53-82.

Palma, José (2023). «Inteligencia artificial y neurociencia. Algunas reflexiones sobre las aportaciones que puede hacer el derecho penal». En Jaime Peris y Antonella Massaro (editores), *Derecho penal, inteligencia artificial y neurociencias* (pp. 249-269). Roma: Roma Tre-Press.

Pérez-Arias, Jacinto (2023). «Algoritmos y big data en la responsabilidad penal: El reto de la cibercriminalidad en el derecho penal». En Jaime Peris y Antonella Massaro (editores), *Derecho penal, inteligencia artificial y neurociencias* (pp. 159-189). Roma: Roma Tre-Press.

Quintero, Gonzalo (2017). «La robótica ante el derecho penal: El vacío de respuesta jurídica a las desviaciones incontroladas». *Revista Electrónica de Estudios Penales y de la Seguridad*, 1: 1-23.

Rettig, Mauricio (2019). *Derecho penal. Parte general.* Tomo 2. 1.ª ed. Santiago: Der.

Rodríguez, Luis (2011). «Naturaleza y fundamento de las circunstancias modificatorias de la responsabilidad criminal». *Revista de Derecho* (Pontificia Universidad Católica de Valparaíso), 36 (1): 397-428.

Romeo, Carlos (2022). «La atribución de responsabilidad penal por los hechos cometidos por sistemas autónomos inteligentes, robótica y tecnologías conexas». *ULP Law Review,* 16 (1-2): 7-16.

—. (2023). «La atribución de responsabilidad penal por los hechos cometidos por sistemas autónomos inteligentes, robótica y tecnologías conexas». En Christian Scheechler (editor), *Los delitos informáticos. Aspectos político-criminales, penales y procesales en la Ley 21.459* (pp. 3-34). Santiago: Der.

Sánchez, Carolina y José Toro-Valencia (2021). «El derecho al control humano: Una respuesta jurídica a la inteligencia artificial». *Revista Chilena de Derecho y Tecnología*, 10 (2): 211-228.

Valls, Javier (2022). «Sobre la responsabilidad penal por la utilización de sistemas inteligentes». *Revista Electrónica de Ciencia Penal y Criminología*, 24-27: 1-35.

Vargas, Tatiana (2011). *Manual de Derecho Penal Práctico. Teoría del delito con casos*. 2.ª ed. Santiago: Abeledo Perrot.

Venezian, Pablo (2022). «Criminal liability of artificial intelligence: The need and opportunity for international regulation». *Centro de Estudios Ius Novum*: 1-5.

Wilenmann, Javier y Miguel Schürmann (2024). «Comentario previo al artículo 1 de la Ley 20.393: La responsabilidad penal corporativa». En Antonio Bascuñán y Javier Wilenmann, *Derecho Penal Económico chileno* (pp. 103-151). Tomo 2. 1.ª ed. Santiago: Der.

Xavier, Túlio (2023). «Inteligencia artificial y responsabilidad penal de personas jurídicas: Un análisis de sus aspectos materiales y procesales». *Estudios Penales y Criminológicos*, 44: 1-39.

Yáñez, Sergio (1994). «La evolución del sistema de derecho penal». *Cuadernos de política criminal*, 54: 1153-1209.

Sobre los autores

Pablo Aguilar Campos es licenciado en Ciencias Jurídicas y Sociales de la Universidad de Chile y magíster en Derecho Penal de la Universidad de Talca/Universitat Pompeu Fabra. Su correo electrónico es pabloaguilarcampos@gmail.com. https://orcid.org/0009-0000-5170-8867.

Víctor Alé Martínez es licenciado en Ciencias Jurídicas y Sociales de la Universidad de Chile y magíster en Derecho Penal por la Universidad de Talca/Universitat Pompeu Fabra. Su correo electrónico es victorale@ug.uchile.cl. https://orcid.org/0009-0007-0501-4910.

REVISTA DE ESTUDIOS DE LA JUSTICIA

La *Revista de Estudios de la Justicia*, fundada en 2002, fue editada inicialmente por el Centro de Estudios de la Justicia hasta 2017. A partir de 2018, su gestión y edición están a cargo del Departamento de Ciencias Penales de la Facultad de Derecho de la Universidad de Chile. Con el propósito de enriquecer el debate jurídico desde perspectivas teóricas y empíricas, la revista ofrece un espacio para difundir el trabajo de académicos de nuestra Facultad, así como de otras casas de estudio nacionales y extranjeras. La *Revista de Estudios de la Justicia* privilegia la publicación de trabajos originales e inéditos sobre temas de interés para las ciencias jurídicas, en cualquiera de sus disciplinas y ciencias afines, con énfasis en investigaciones relacionadas con reformas a la justicia.

DIRECTOR

Álvaro Castro

(acastro@derecho.uchile.cl)

SITIO WEB

rej.uchile.cl

CORREO ELECTRÓNICO

rej@derecho.uchile.cl

La edición de textos, el diseño editorial
y la conversión a formatos electrónicos de este artículo
estuvieron a cargo de Tipográfica
(www.tipografica.io)