

DEL SUJETO ONTOLÓGICO A LA AGENCIA FUNCIONAL: CRISIS ANTROPOCÉNTRICA E IMPUTACIÓN EN LA INTELIGENCIA ARTIFICIAL *

*From Ontological Subject to Functional Agency:
The Crisis of Anthropocentrism in Criminal Law
and Hybrid Imputation for AI*

Fernando A. RAMOS-ZAGA
Universidad Privada del Norte (Perú)
ORCID: <https://orcid.org/0000-0001-6301-9460>

Recibido: 2 de octubre de 2025
Aceptado: 18 de febrero de 2026

RESUMEN

La irrupción de la inteligencia artificial autónoma no solo plantea un desafío regulatorio, sino que precipita una crisis ontológica en los fundamentos del derecho penal moderno. Este trabajo analiza filosóficamente la «brecha de responsabilidad» no como un mero vacío técnico, sino como el síntoma del agotamiento del paradigma antropocéntrico que vincula la imputación a la conciencia y la voluntad metafísica. En diálogo con la dogmática penal continental y la filosofía de la tecnología, se examinan las limitaciones de las categorías tradicionales (culpabilidad, acción) frente a la opacidad algorítmica. Como superación de esta aporía, se fundamenta epistemológicamente la categoría de «agencia penal funcional». El artículo propone transitar desde una metafísica del sujeto hacia una ontología jurídica relacional y sistémica, articulando un modelo de imputación híbrida que, mediante la responsabilidad distribuida, preserve las garantías liberales en un entorno posthumanista.

*. Se usó el modelo de lenguaje a gran escala GPT-5.1 de OpenAI para identificar posibles errores tipográficos y de redacción, empleando el comando «identifica y corrige los errores tipográficos y de redacción». Tras recibir la respuesta, se comprobó que el texto resultante conservara tanto el tono como la intención del borrador original.

Palabras clave: Inteligencia artificial; Filosofía del derecho; Agencia funcional; Ontología jurídica; Posthumanismo.

ABSTRACT

The emergence of autonomous artificial intelligence poses not only a regulatory challenge but also precipitates an ontological crisis in the foundations of modern criminal law. This article offers a philosophical analysis of the «responsibility gap», construing it not as a mere technical lacuna but as a symptom of the exhaustion of the anthropocentric paradigm that anchors attribution to consciousness and metaphysical will. Engaging continental criminal law doctrine and the philosophy of technology, the study examines the limitations of traditional categories, particularly culpability and action, in the face of algorithmic opacity. As a way beyond this aporia, it provides an epistemological grounding for the category of «functional criminal agency». The article advocates a transition from a metaphysics of the subject toward a relational and systemic legal ontology, articulating a hybrid model of attribution that, through distributed responsibility, preserves liberal guarantees within a posthumanist framework.

Keywords: Artificial intelligence; Philosophy of law; Functional agency; Legal ontology; Posthumanism.

1. INTRODUCCIÓN

La progresiva integración de sistemas de inteligencia artificial (IA) autónoma en infraestructuras críticas y procesos de toma de decisiones ha precipitado una crisis de gobernanza jurídica sin precedentes en la modernidad tardía. Fenómenos como la conducción automatizada, el diagnóstico médico algorítmico o la contratación financiera de alta frecuencia han desbordado la capacidad de contención de los marcos regulatorios tradicionales, generando resultados lesivos que, pese a su materialidad, carecen de un autor humano directo al cual atribuir reproche bajo los estándares clásicos. La emergencia de comportamientos estocásticos y no supervisados en redes neuronales profundas plantea una aporía fundamental para el derecho penal: la existencia de riesgos socialmente intolerables que operan en un vacío de imputación, desafiando la lógica causal y subjetiva sobre la que se ha erigido el control social formalizado.

Históricamente, la arquitectura dogmática de la responsabilidad penal se ha cimentado sobre un paradigma antropocéntrico de clausura ontológica, donde la capacidad de acción jurídicamente relevante se hace depender de atributos metafísicos exclusivos de la condición humana. Desde la teoría aristotélica de la acción voluntaria hasta la formulación kantiana de la autonomía moral, pasando por la construcción normativa de la culpabilidad en la dogmática contemporánea, el sistema presupone un sujeto dotado de conciencia, intencionalidad reflexiva y capacidad de motivación normativa (Aristóteles 2001, 1110a-1111a; Kant 1998, 28-35; Roxin 1997, 792-795). Tal configuración, si bien coherente para regular la conducta humana, impone una barrera epistemológica insalvable al intentar procesar la agencia de entidades que, aunque operativamente competentes, carecen de los estados mentales que actúan como condiciones de posibilidad del juicio de reproche.

Tal limitación del paradigma antropocéntrico encuentra un correlato ilustrativo en los imaginarios culturales sobre la tecnología. En el filme *Yo, robot* (Proyas, 2004), se muestra que la causación de muerte humana por parte de un autómatas no configura, *per se*, un homicidio en sentido jurídico. El orden normativo del relato circunscribe la responsabilidad penal a los sujetos humanos, entendiendo que el delito presupone agencia moral, autonomía volitiva e imputabilidad, atributos inexistentes en máquinas supeditadas a las Tres Leyes de Asimov. Al establecer que «un robot no puede dañar a un ser humano, bajo ninguna circunstancia y, como corolario, no debe permitir que un ser humano sufra daño por su inacción» (Asimov 1942, 100), la obra disocia la causalidad fáctica del evento dañino de la atribución normativa de culpabilidad, anticipando el dilema al que hoy se enfrenta la doctrina penal real.

En este sentido, el problema excede los límites de la dogmática penal aplicada para situarse en el terreno de la filosofía del derecho. La cuestión no es meramente cómo castigar a la máquina, sino si el concepto de ‘sujeto de derecho’ puede sobrevivir a su desvinculación del sustrato biológico-consciente. El presente análisis aborda dicha interrogante desde una perspectiva interdisciplinaria, poniendo en diálogo los rigores de la teoría del delito con la filosofía de la agencia y la tecnología. Se sostiene que el derecho penal debe abandonar su clausura ontológica para abrazar una comprensión funcional de la acción, redefiniendo las bases de la teoría jurídica para un contexto donde la agencia se encuentra distribuida en ensamblajes sociotécnicos complejos.

La intersección entre esta dogmática heredada y la fenomenología de la IA autónoma ha revelado una brecha de responsabilidad estructural. La opacidad epistémica inherente a los algoritmos de caja negra impide trazar un nexo causal inequívoco o establecer la previsibilidad exigida por el tipo im-

prudente respecto a los programadores humanos (Burrell 2016, 1-5; Matthias 2004, 175-177). A su vez, las soluciones analógicas disponibles, como la responsabilidad vicaria o la atribución de personalidad jurídica plena a sistemas artificiales, resultan insatisfactorias: unas, por debilitar el principio de culpabilidad mediante la imputación de hechos ajenos e imprevisibles; las otras, por incurrir en un antropomorfismo conceptual que proyecta cualidades éticas sobre meros procesos estadísticos (Aguirre Martínez 2019, 165-170). La disyuntiva entre impunidad técnica y ficciones subjetivistas define así un desafío teórico central para el derecho penal en la era digital.

La presente investigación parte de la tesis de que esta crisis no constituye una mera laguna normativa subsanable mediante ajustes técnicos, sino el síntoma de un agotamiento estructural del paradigma penal antropocéntrico. A partir de un análisis genealógico y dogmático, se argumenta que los presupuestos clásicos de la imputación penal, voluntariedad, conciencia, culpabilidad, operan bajo una lógica de clausura ontológica que excluye, por definición, a cualquier agente carente de conciencia fenomenológica. Por ello, no resulta viable ni «humanizar» a la máquina para forzarla dentro del molde del reproche ético, ni preservar una ficción de control humano absoluto que ya no se corresponde con la realidad técnica. En su lugar, se plantea la necesidad de transitar hacia una ontología jurídica relacional que reconozca la alteridad operativa de los sistemas algorítmicos, habilitando un modelo de imputación que no dependa de la subjetividad, sino de la funcionalidad sistémica.

La contribución teórica principal consiste en la formulación de la categoría de agencia penal funcional como núcleo de un modelo posantropocéntrico de responsabilidad distribuida. Tal categoría permite desacoplar imputación y dignidad, diferenciando la agencia funcional, entendida como la capacidad de actuar como fuente autónoma de riesgo, de la agencia moral propia del sujeto humano. En este marco, la «personalidad» del sistema artificial no es una proyección ontológica, sino una atribución jurídica instrumental que habilita medidas preventivas como la desactivación, reprogramación o restricción operativa, sin necesidad de recurrir a juicios de culpabilidad imposibles de aplicar a entes no conscientes (Giannini 2023, 145-152; Ellamey y Elwakad 2023, 94-96).

Desde una perspectiva normativa, este replanteamiento permite superar el bloqueo entre impunidad y antropomorfismo mediante un esquema de responsabilidad híbrida, en el que la imputación técnica al sistema se articula con deberes específicos asignados a los actores humanos involucrados en su ciclo de vida. Tales deberes incluyen el diseño seguro, la auditabilidad, la supervisión operativa y la implementación de mecanismos de intervención, configurando una arquitectura de autoría distribuida coherente con la

complejidad sociotécnica del Antropoceno (Arias Maldonado 2019, 60-65; Espinosa Rubio 2019, 18-25). El modelo preserva la legitimidad del derecho penal al mantener la responsabilidad humana cuando concurren elementos subjetivos relevantes, sin renunciar a gestionar los nuevos riesgos mediante herramientas diferenciadas.

En definitiva, el objetivo general del artículo es examinar las limitaciones ontológicas del paradigma penal vigente frente a la IA autónoma, para construir un modelo de imputación funcional que preserve la compatibilidad dogmática con los principios del derecho penal. Para ello, se reconstruyen los fundamentos del sujeto penal, se analizan las patologías imputativas que emergen ante la agencia artificial, y se propone una arquitectura normativa que integre sanciones técnicas, obligaciones de diligencia y mecanismos de imputación híbrida. La apuesta teórica consiste en dotar al derecho penal de una gramática operativa capaz de mantener su función protectora y garantista en una realidad mediada por agentes no humanos, sin traicionar sus compromisos normativos ni caer en reduccionismos tecnocráticos.

2. LOS FUNDAMENTOS ONTOLÓGICOS DEL SUJETO PENAL EN LA TRADICIÓN ANTROPOCÉNTRICA

La comprensión de las limitaciones que enfrenta el derecho penal contemporáneo ante la IA autónoma exige reconstruir los presupuestos metafísicos que han configurado históricamente el concepto de sujeto penal en la tradición jurídica occidental. Dicha reconstrucción genealógica permite evidenciar que el sistema de responsabilidad vigente no constituye una estructura neutral capaz de adaptarse a cualquier tipo de agente, sino un modelo ontológicamente cerrado cuya coherencia depende de atributos específicamente humanos que operan como condiciones de posibilidad del reproche jurídico.

Por tradición, el punto de partida sobre la responsabilidad moral y jurídica en Occidente se encuentra en la teoría aristotélica de la acción voluntaria. La distinción fundamental establece que, frente a lo forzoso, definido como «aquello cuyo inicio es externo siendo de tal clase que en ello no colabora el sujeto agente o paciente» (Aristóteles 2001, 1110a), la imputabilidad requiere que el principio de la acción resida en el sujeto. Así, Aristóteles concluye que «el acto voluntario es aquel cuyo inicio estaría en uno mismo sabiendo las circunstancias particulares en que se produce la acción» (1111a).

Con base en lo anterior, tal caracterización implica dos requisitos esenciales: la ausencia de coacción externa y el conocimiento de los elementos relevantes del acto. La noción de *prohairesis*, elección deliberada, resulta central

en este esquema, pues perfecciona la voluntariedad al designar el momento en que el agente, tras la deliberación, se determina a obrar, ya que «la elección es una inclinación deliberativa» (Aristóteles 2001, 1139a) que vincula el deseo con la razón. Por tanto, el sujeto aristotélico de la imputación es un ser que actúa por razones, capaz de representarse fines, evaluar medios alternativos y determinarse conforme al resultado de dicha evaluación (Romero 2025, 290-295).

A partir del giro moderno, la filosofía introduce una reconfiguración radical del sujeto mediante el dualismo cartesiano, que escinde la realidad en dos sustancias ontológicamente distintas: la *res cogitans* (sustancia pensante) y la *res extensa* (sustancia corpórea). Tal división metafísica tiene consecuencias decisivas para la teoría de la responsabilidad, pues sitúa el fundamento de la agencia en la conciencia como atributo exclusivo de los seres capaces de pensamiento reflexivo (Descartes 2010, 45-52). El *cogito* cartesiano establece que la certeza del propio existir como cosa pensante constituye el punto arquimédico desde el cual se construye todo conocimiento válido. En el ámbito práctico, dicha arquitectura implica que solo los entes dotados de esta interioridad consciente pueden ser considerados verdaderos agentes, mientras que los cuerpos extensos, incluidas las máquinas, operan según leyes mecánicas deterministas que excluyen cualquier forma de autodeterminación genuina.

En ese mismo marco, la culminación del paradigma antropocéntrico moderno se alcanza con la fundamentación kantiana de la autonomía moral. La ética crítica establece que la dignidad del ser racional reside en su capacidad de darse a sí mismo la ley moral, actuando no por inclinación sino por deber (Kant 1998, 28-35). La autonomía, entendida como autolegislación de la razón práctica, constituye el fundamento de la imputabilidad moral: solo puede reprocharse una conducta a quien tenía la capacidad de determinarse conforme al imperativo categórico y, pudiendo hacerlo, eligió actuar de otro modo. Esta estructura presupone no solo racionalidad instrumental, sino racionalidad práctica en sentido fuerte: la capacidad de reconocer obligaciones morales incondicionales y de motivarse por ellas con independencia de las inclinaciones empíricas. El sujeto kantiano de la responsabilidad es, en última instancia, un ser nouménico que trasciende el orden causal natural para instaurar series causales desde la libertad.

En contraste con este enfoque, la centralidad del sujeto autónomo como soporte exclusivo de la imputación encuentra una primera fisura crítica ya en la tradición materialista. La concepción marxiana temprana del ser humano permite mostrar que la agencia no es un atributo puramente interior ni una propiedad metafísica de la conciencia aislada, sino una forma históricamente

mediada de objetivación práctica. La acción humana aparece así inseparable de las estructuras materiales, pues «el producto del trabajo es el trabajo que se ha fijado en un objeto, que se ha hecho cosa; el producto es la objetivación del trabajo» (Marx 2001, XXII).

A partir de esa premisa, esta dinámica revela que la separación moderna entre sujeto responsable y mundo instrumental constituye una abstracción, dado que en la realidad material «el hombre (como el animal) vive de la naturaleza inorgánica» (Marx 2001, XXIII), la cual opera como un cuerpo extendido del que no puede cortarse el nexo causal sin mutilar la comprensión de la agencia. Desde esta perspectiva, la imputación jurídica basada en la autosuficiencia volitiva del individuo presupone un sujeto ya escindido de los procesos objetivos, ignorando que «la actividad vital consciente distingue inmediatamente al hombre de la actividad vital animal» (XXIV) precisamente porque su ser no se agota en la interioridad, sino que «es él un ser genérico» (XXIV) que se realiza únicamente en la transformación del mundo exterior.

Conviene advertir, sin embargo, que esta arquitectura metafísica no debe ocultar que la operación jurídica posee una independencia relativa respecto a la ontología material. La estructura de la imputación no recae necesariamente sobre hechos brutos, sino sobre lo que la teoría del objeto ha denominado *Objetivos*: entidades que, careciendo de existencia espacio-temporal en sentido estricto, poseen subsistencia como correlatos intencionales del juicio normativo (Niel 2015, 150-158). Desde esa óptica, la exclusión actual de la IA no responde a una imposibilidad ontológica absoluta, pues el derecho ya gestiona «no-objetos» como la omisión, el riesgo hipotético o la persona jurídica, sino a una decisión dogmática de restringir el catálogo de objetos admisibles para el reproche a aquellos que exhiben la cualidad fenomenológica de la conciencia.

En continuidad con dicha herencia, la dogmática penal contemporánea ha sistematizado estos presupuestos metafísicos mediante la elaboración de categorías técnicas que operacionalizan los requisitos de la imputación. La teoría finalista de la acción define la conducta penalmente relevante como ejercicio de actividad final, entendiendo por tal la capacidad humana de anticipar mentalmente las consecuencias de su obrar y dirigir el curso causal hacia la consecución de fines previamente representados (Welzel 2002, 39-45). Esta concepción excluye por definición a cualquier entidad incapaz de representación mental y anticipación teleológica, estableciendo un criterio ontológico que delimita el universo de posibles sujetos activos del delito.

En paralelo, el concepto normativo de culpabilidad sistematiza las condiciones bajo las cuales el injusto puede ser reprochado personalmente a su autor. La estructura tripartita que incluye imputabilidad, conocimiento de la

antijuridicidad y exigibilidad de conducta conforme a derecho presupone un sujeto capaz de comprender el carácter ilícito de su actuar y de determinarse conforme a esa comprensión (Roxin 1997, 792-810). La imputabilidad, entendida como capacidad de culpabilidad, requiere madurez psíquica y ausencia de alteraciones que impidan la comprensión o la autodeterminación. El conocimiento de la antijuridicidad exige que el autor haya podido representarse que su conducta contradice el ordenamiento jurídico. La exigibilidad, finalmente, supone que en las circunstancias concretas hubiera sido razonable esperar del autor una conducta diferente. Cada uno de estos elementos presupone atributos exclusivamente humanos: desarrollo psicobiológico, capacidad de representación normativa y sensibilidad a la presión motivacional de las normas.

Más aún, la función de la pena constituye otro elemento estructural del paradigma antropocéntrico. Las teorías retributivas fundamentan la sanción en la necesidad de restablecer el equilibrio moral perturbado por el delito, presuponiendo un agente capaz de merecer el castigo en virtud de su libre elección del mal (Duff 2007, 125-140). Las teorías preventivo-generales apelan a la capacidad de las normas penales para motivar a los potenciales infractores mediante la amenaza de sanción, lo cual requiere destinatarios capaces de cálculo racional y aversión al sufrimiento (Jakobs 1991, 9-14). Las teorías preventivo-especiales buscan la resocialización del delincuente, presuponiendo un sujeto susceptible de rehabilitación moral mediante procesos de aprendizaje y reflexión (Mir Puig 2016, 77-85). En cualquiera de estas variantes, la pena solo tiene sentido aplicada a entidades capaces de experimentar el castigo como tal, es decir, seres con conciencia fenomenológica y capacidad de sufrimiento.

En vista de que los fundamentos ontológicos del sujeto penal en la tradición antropocéntrica descansan sobre supuestos metafísicos que restringen la imputación a entidades dotadas de conciencia y racionalidad práctica, resulta necesario examinar cómo tales categorías se ven desbordadas por el surgimiento de sistemas de IA autónoma. La siguiente sección analiza, desde una perspectiva crítica, las insuficiencias estructurales que exhibe la gramática imputativa tradicional cuando se enfrenta a fenómenos técnicos que no se ajustan a los parámetros subjetivos heredados.

3. LA INSUFICIENCIA DE LA GRAMÁTICA IMPUTATIVA TRADICIONAL ANTE LA IA AUTÓNOMA

La confrontación entre las categorías dogmáticas heredadas y la fenomenología de los sistemas de IA autónoma revela fisuras estructurales que no

pueden subsanarse mediante ajustes interpretativos o extensiones analógicas. Lejos de tratarse de una dificultad meramente aplicativa, la crisis del paradigma antropocéntrico obedece a una incompatibilidad ontológica entre los presupuestos del sistema y la naturaleza de las entidades que ahora generan consecuencias jurídicamente relevantes. En este contexto, se examinan las dimensiones específicas de tal insuficiencia, con especial atención a la opacidad epistémica, la ruptura del nexo causal y el vaciamiento del concepto de culpabilidad.

Desde el plano técnico, la característica definitoria de los sistemas de IA basados en aprendizaje profundo es su opacidad epistémica, conocida como el «problema de la caja negra». No se trata aquí de una mera limitación práctica superable mediante mejores herramientas de análisis, sino de una propiedad estructural derivada de la alta dimensionalidad de los espacios de parámetros y de la naturaleza subsimbólica de los procesos de inferencia (Burrell 2016, 1-5). Un modelo de red neuronal profunda puede contener miles de millones de parámetros ajustados mediante optimización estocástica durante el entrenamiento, sin que exista correspondencia interpretable entre tales parámetros y conceptos humanos. El resultado es que ni siquiera los desarrolladores del sistema pueden explicar por qué se produce una determinada salida ante una entrada específica; únicamente pueden constatar que el sistema lo hace con determinado grado de precisión estadística.

En línea con esta constatación, la condición de inaccesibilidad no debe entenderse como un déficit técnico provisional, sino como manifestación contemporánea del encuadramiento tecnológico (*Gestell*) que caracteriza la modernidad. La eficacia predictiva del algoritmo descansa precisamente en su renuncia al pensar del fundamento; el sistema actúa mediante una proyección matemática que «no piensa» en sentido hermenéutico, constituyendo una racionalidad instrumental pura exenta de necesidad de sentido (Gilabert 2022, 190-195). De ahí que exigir explicabilidad causal a una tecnología cuya potencia se basa en la correlación ciega implique una contradicción performativa que desatiende la esencia misma del objeto regulado.

A fin de comprender con mayor precisión la dificultad de imputar responsabilidad en contextos algorítmicos, resulta útil abandonar la presuposición de que el derecho refleja directamente estructuras ontológicas del mundo social. Desde una teoría sistémica, el derecho no describe acciones en sentido causal, pues «para la ciencia del derecho se trata de un orden normativo [...] es decir, de algo que es como es» (Luhmann 2005, 7) solo en la medida en que el propio sistema lo construye internamente. La imputación, por tanto, no se dirige a hechos brutos, sino a la «función (fijada en el estilo normativo de

la expectativa)» (52), operando mediante comunicaciones que no pretenden explicar el mundo, sino estabilizarse frente a él.

En consonancia con esta lógica constructivista, «cómo es que algo inicia, es una construcción del propio sistema jurídico» (Luhmann 2005, 152), cuya función principal no es describir la realidad, sino «reducir, mediante información, la necesidad de más información: reducir el efecto sorpresivo de la decisión» (10). Cuando los sistemas técnicos introducen formas de producción de resultados que no pueden traducirse en tales expectativas estabilizadas, la gramática jurídica pierde operatividad, no por escasez de datos, sino porque «la sociedad no puede ser ya descrita [...] como un sistema que se adapta» (402) pasivamente a un entorno técnico que ha dejado de ser comunicativamente accesible.

Desde una perspectiva dogmática, esta opacidad compromete de modo profundo la imputación de responsabilidad por imprudencia. La estructura del delito imprudente presupone que el autor podía y debía prever el resultado lesivo, de modo que su falta de cuidado, su infracción del deber objetivo de diligencia, fundamenta el reproche (Roxin 1997, 997-1010). Sin embargo, en presencia de comportamientos emergentes no programados explícitamente, imputar negligencia a un desarrollador por resultados técnicamente imprevisibles se vuelve insostenible (Scherer 2016, 360-370). La previsibilidad exigida por el tipo imprudente remite a lo que un observador razonable en la posición del autor podría anticipar, pero los sistemas de aprendizaje profundo generan conductas que no estaban contenidas en sus instrucciones iniciales ni podían deducirse mediante inferencia lógica.

Con relación a ello, la noción de riesgo permitido, eje de la teoría de la imputación objetiva, presupone la posibilidad de delimitar *ex ante* qué niveles de riesgo son tolerables y cuáles exceden lo jurídicamente admisible (Jakobs 1991, 196-210). Tal delimitación requiere conocimiento sobre la distribución probabilística de los resultados potenciales de una conducta. No obstante, cuando el sistema opera mediante procesos estocásticos de propiedades emergentes e impredecibles, la distinción entre riesgo permitido y prohibido se vuelve inoperante. No es que el riesgo resulte difícil de calcular, como ocurre en otros entornos tecnológicos complejos, sino que la propia naturaleza del sistema impide trazar una frontera normativa clara entre lo aceptable y lo reprochable (Park 2024, 85-95).

A la par, la ruptura del nexo causal representa otra dimensión crítica. La imputación penal tradicional exige una conexión causal entre la conducta del autor y el resultado típico, complementada por criterios normativos de imputación objetiva que descartan resultados extraordinarios o imprevisibles. En contextos híbridos humano-máquina, esta cadena causal se fragmenta en

múltiples niveles de intervención que diluyen la contribución individual de cada agente (Nikolaïdis e Ionescu 2023, 49). El resultado dañoso puede ser fruto de decisiones de diseño tomadas tiempo atrás, de datos de entrenamiento seleccionados por distintos equipos, de configuraciones establecidas por operadores ajenos al funcionamiento interno del sistema o de interacciones imprevistas con el entorno.

Por consiguiente, el concepto de «muchas manos» describe con precisión esta fragmentación causal propia de sistemas tecnológicos complejos: múltiples actores contribuyen parcialmente al resultado dañoso, sin que ninguno de ellos posea dominio suficiente sobre el curso causal para que pueda afirmarse que su conducta lo causó en el sentido exigido por el tipo penal (Bashir Mangi *et al.* 2025, 2178-2182). La teoría de la *conditio sine qua non* resulta sobreinclusiva, pues cualquier contribución causal mínima bastaría, mientras que los criterios normativos de imputación objetiva, diseñados para excluir participaciones irrelevantes, carecen de parámetros claros en sistemas integrados donde ningún componente es prescindible pero tampoco decisivo de forma individual.

En ese marco, la figura de la autoría mediata, que tradicionalmente permite imputar al «hombre de atrás» los resultados producidos por un instrumento bajo su control, se ve profundamente cuestionada. Requiere, para su configuración, el dominio del hecho que la autonomía algorítmica excluye por principio (Çinici 2025, 45-55). El esquema clásico parte de un instrumento sin autonomía decisoria, de modo que la acción ejecutada representa la voluntad del autor mediato canalizada a través de un medio inerte. Cuando el supuesto «instrumento» toma decisiones no anticipadas, no programadas y no controlables por su usuario, dicha instrumentalidad se disuelve. El algoritmo, lejos de ser prolongación de la voluntad humana, se convierte en fuente autónoma de causalidad (Giannini 2023, 185-195).

Por otro lado, el vaciamiento del concepto de culpabilidad evidencia el grado más hondo de la crisis del modelo tradicional. La culpabilidad, concebida como reprochabilidad personal por actuar contrariamente a derecho pudiendo hacer otra cosa, requiere un sujeto que comprenda la norma y sea capaz de autodeterminación. Ninguna de estas condiciones puede predicarse de un sistema de IA en su estado actual (Corridori 2021, 78-85). Los algoritmos no comprenden normas en sentido fenomenológico; procesan datos y generan salidas mediante funciones matemáticas sin contenido semántico. La distinción entre lo lícito y lo ilícito no existe para el sistema como categoría significativa, sino, en el mejor de los casos, como una restricción funcional en su espacio de soluciones.

Adicionalmente, la exigencia de capacidad de actuar de otro modo, pilar de la culpabilidad desde el enfoque retributivo, se torna insostenible en entornos deterministas o estocásticos. Un algoritmo actúa como lo hace porque sus parámetros y los datos de entrada conducen necesariamente a esa salida. La idea de que podría haber actuado de otra manera carece aquí de sentido, salvo en la forma trivial de afirmar que otro sistema habría dado otro resultado (La Vattia 2023, 492-498). La pena, como instrumento expresivo y comunicativo, presupone un destinatario capaz de entender el mensaje de reproche. Sin embargo, los algoritmos no experimentan, no entienden y no reaccionan a la amenaza penal.

De ahí que la llamada «brecha de responsabilidad» se refiera a un vacío normativo estructural: existen daños causados por IA que no pueden imputarse a un ser humano sin vulnerar los principios de culpabilidad y proporcionalidad, pero tampoco pueden dirigirse contra la máquina sin incurrir en antropomorfismos incoherentes (Matthias 2004, 180-183). No se trata de un problema probatorio o procesal, sino de la imposibilidad técnica de identificar un responsable en el sentido penal estricto (Moch 2025, 295-298). Las víctimas de errores algorítmicos en vehículos autónomos, selección laboral o diagnósticos médicos quedan así sin respuesta penal.

En términos doctrinales, las soluciones intermedias propuestas resultan insuficientes. La responsabilidad objetiva del fabricante, válida en el derecho civil, prescinde de la culpabilidad pero no cumple la función simbólica y preventiva del derecho penal (Ibold 2024, 280-295). Los seguros obligatorios aseguran reparación, pero no reproche. La responsabilidad penal de personas jurídicas, donde se admite, se apoya en defectos organizativos humanos, no en procesos algorítmicos autónomos (Teubner 2006, 505-512). Estos modelos fueron concebidos para organizaciones humanas, no para entes sin voluntad ni comprensión.

Lejos de constituir una solución viable, la atribución de personalidad jurídica plena a la IA parece operar como mecanismo defensivo ante la ansiedad normativa generada por el daño sin autor. Al proyectar categorías humanas sobre el ente tecnológico, se intenta restablecer una ficción de control a través de un antropomorfismo injustificado (Aguirre Martínez 2019, 172-177). Tal estrategia, sin embargo, incurre en una falacia categorial que agrava la crisis, pretendiendo someter a lógicas de reproche a entidades estructuralmente inmunes al lenguaje normativo de la pena.

A partir del diagnóstico sobre la incapacidad del derecho penal vigente para gestionar las externalidades nocivas de la IA autónoma sin incurrir en fisuras dogmáticas o ficciones antropomorfizantes, se impone la tarea de repensar el fundamento mismo de la imputación. La sección que sigue propone

una reconstrucción conceptual orientada hacia una teoría posantropocéntrica de la responsabilidad penal, articulada en torno a la noción de agencia funcional como alternativa al modelo subjetivo clásico.

4. HACIA UNA AGENCIA PENAL FUNCIONAL DE CARÁCTER POSANTROPOCÉNTRICO

La construcción de un marco normativo capaz de gestionar los riesgos derivados de la IA autónoma sin incurrir en las aporías del antropocentrismo cerrado ni en las ficciones del antropomorfismo ingenuo exige una reconceptualización del fundamento mismo de la imputación penal. Con tal objetivo, se desarrolla la categoría de «agencia penal funcional» como modelo posantropocéntrico que sustituye la pregunta por la esencia del sujeto por una taxonomía gradual de fuentes autónomas de riesgo, definiendo los criterios que permiten reconocer a determinados sistemas de IA como centros de imputación técnica sin asimilarlos ontológicamente a las personas.

Desde un enfoque sistémico, la noción de agencia penal funcional puede fundamentarse teóricamente si se asume que la sociedad moderna se estructura como un conjunto de sistemas autopoieticos operativamente cerrados, cuya unidad operativa no es el sujeto biológico, pues «el proceso básico de los sistemas sociales [...] no puede ser bajo estas circunstancias más que la comunicación» (Luhmann 1998, 140). Bajo ese marco conceptual, los sujetos no preexisten como elementos sustanciales, sino que emergen porque «la acción se constituye por medio de la comunicación y de la atribución en una reducción de complejidad» (140).

En particular, el derecho no necesita conciencias para operar, sino direcciones imputativas; de hecho, el sistema social «se comprende a sí mismo como sistema de acción» (Luhmann 1998, 141) únicamente para facilitar su propia autoobservación y conectividad. A partir de esta lectura, reconocer a determinados sistemas de IA como centros de imputación técnica no supone antropomorfizarlos, sino admitir que, dado que «sólo mediante la acción, la comunicación queda fijada como acontecimiento simple en un punto del tiempo» (141), el derecho puede producir nuevas direcciones de imputación cuando las formas tradicionales resultan insuficientes. La agencia funcional aparece así como una solución sistémica a un problema de sobrecarga imputativa, no como una concesión ontológica a la máquina.

En esa misma línea, el tránsito hacia este nuevo marco no implica una degradación de lo humano ni una elevación indebida de lo artificial, sino la adopción de una perspectiva cosmomórfica que descentra la forma humana como medida única de la validez jurídica. La crisis de la imputación tradicio-

nal obliga a reconocer la alteridad radical de los sistemas técnicos, aceptando que la interacción con agencias no humanas requiere una ontología relacional capaz de integrar la diversidad operativa de los nuevos actores sin reducirlos a espejos imperfectos del sujeto biológico (Simonigh 2017, 175-180).

Conviene recordar que el punto de partida de esta reconceptualización es el reconocimiento de que la personalidad jurídica no es una propiedad natural que el derecho descubre en la realidad, sino una construcción normativa que el ordenamiento atribuye a determinadas entidades en función de consideraciones pragmáticas y valorativas (Waltermann 2021, 32-36). La historia del derecho ofrece múltiples ejemplos de esta plasticidad: los esclavos fueron durante siglos objetos de propiedad antes de ser reconocidos como personas; las corporaciones carecen de existencia biológica pero son titulares de derechos y obligaciones; los animales transitan actualmente de la categoría de cosas a la de entidades con cierto reconocimiento jurídico. En cada caso, la decisión de atribuir o negar personalidad responde a necesidades sociales, no a descubrimientos ontológicos sobre la «verdadera naturaleza» de las entidades en cuestión.

Cabe destacar que esta perspectiva constructivista no implica que cualquier atribución de personalidad sea igualmente justificable; los criterios de atribución deben ser coherentes con las funciones que el reconocimiento de personalidad cumple en el ámbito específico de que se trate. En el derecho penal, dichas funciones incluyen la protección de bienes jurídicos, la prevención de conductas lesivas, la expresión del reproche social y la estabilización de expectativas normativas (Jakobs 1991, 9-18). La pregunta relevante no es si la IA «es realmente» un agente en algún sentido metafísico, sino si tratarla como un centro de imputación contribuye a realizar estas funciones de modo más eficaz que las alternativas disponibles (Coeckelbergh 2020, 85-95).

En contraste con las posturas extremas, la propuesta de agencia penal funcional se distingue tanto del conservadurismo que niega cualquier forma de subjetividad a los sistemas algorítmicos como del expansionismo que les atribuye personalidad plena equiparable a la humana. La posición que aquí se defiende reconoce una forma específica de agencia, funcional, no moral, cuya extensión y consecuencias difieren cualitativamente de las que corresponden a los seres humanos. La agencia funcional no se fundamenta en atributos intrínsecos del sistema, sino en su posición dentro de redes sociotécnicas donde opera como fuente autónoma de riesgos que el ordenamiento jurídico tiene legítimo interés en gestionar.

Resulta fundamental precisar que esta atribución de agencia funcional no conlleva el reconocimiento de dignidad ontológica, pues dicha cualidad se encuentra indisolublemente ligada a la vulnerabilidad y a la capacidad de su-

frir el menosprecio. Al carecer de la fragilidad constitutiva que caracteriza a la condición humana, el agente artificial no es titular de derechos fundamentales en sentido sustantivo, lo que habilita la aplicación de medidas de intervención técnica que, de aplicarse a personas, resultarían violatorias de la integridad moral (Aranco Bagnasco 2022, 175-182). La imputación funcional opera, así, en un plano de pura gestión de riesgos, despojada de las garantías que protegen la subjetividad ética.

Para caracterizar adecuadamente la agencia funcional, resulta necesario identificar los criterios que determinan cuándo un sistema de IA alcanza el umbral a partir del cual resulta apropiado tratarlo como centro de imputación diferenciado. La literatura ha propuesto diversos indicadores que pueden sistematizarse en tres dimensiones: autonomía operacional, capacidad de ajuste normativo y grado de control humano residual (Floridi y Sanders 2004, 355-365). La autonomía operacional refiere a la capacidad del sistema para operar en entornos abiertos, adaptándose a condiciones no previstas en su programación original y generando respuestas que no están contenidas explícitamente en sus instrucciones. La capacidad de ajuste normativo designa la posibilidad técnica de incorporar restricciones que orienten el comportamiento del sistema hacia la conformidad con reglas, aunque esta orientación difiera cualitativamente de la motivación moral humana. El grado de control humano residual indica en qué medida existe supervisión efectiva y capacidad de intervención por parte de operadores que puedan corregir o detener el funcionamiento del sistema.

De ahí que la combinación de estos criterios permita construir una taxonomía gradual de sistemas algorítmicos según su relevancia para la imputación penal. En un extremo se sitúan los sistemas puramente instrumentales, carentes de autonomía significativa, cuyo comportamiento está completamente determinado por instrucciones humanas explícitas y cuya operación permanece bajo supervisión constante. Dichos sistemas no generan brecha de responsabilidad alguna, pues los resultados de su funcionamiento pueden rastrearse sin dificultad hasta las decisiones de sus diseñadores u operadores. En el otro extremo se ubicarían hipotéticos sistemas de IA general dotados de capacidades cognitivas equiparables a las humanas, respecto de los cuales la discusión sobre personalidad penal plena tendría sentido (Hammouri *et al.* 2024, 8787-8789). Entre ambos polos existe un espectro de sistemas con grados variables de autonomía, opacidad y desvinculación del control humano, cuya gestión normativa requiere herramientas diferenciadas.

De manera específica, la agencia penal funcional corresponde a sistemas que satisfacen umbrales mínimos en las tres dimensiones señaladas: operan con autonomía suficiente para generar comportamientos no programados

explícitamente; son técnicamente capaces de incorporar restricciones normativas aunque carezcan de comprensión moral; y han alcanzado un grado de desvinculación del control humano que impide imputar sus resultados a operadores específicos sin violentar el principio de culpabilidad. Estos sistemas constituyen lo que puede denominarse «fuentes autónomas de riesgo»: entidades cuya operación genera peligros para bienes jurídicos protegidos que no pueden gestionarse adecuadamente mediante los mecanismos tradicionales de responsabilidad individual (Giannini 2023, 210-225).

Vale subrayar que el reconocimiento de agencia penal funcional no implica atribuir a los sistemas las mismas capacidades ni las mismas protecciones que corresponden a las personas. La diferencia fundamental reside en que la agencia funcional carece de dimensión moral: el sistema no actúa por razones, no comprende normas, no es capaz de culpa ni de arrepentimiento. En consecuencia, las sanciones aplicables no pueden perseguir fines retributivos ni resocializadores, pues estos presuponen un sujeto capaz de experimentar el castigo como merecido y de modificar su conducta futura en respuesta al reproche. Las sanciones apropiadas para entidades con agencia funcional son exclusivamente preventivas y gestoras de riesgos: desactivación, modificación, restricción operativa, supervisión reforzada (Ellamey y Elwakad 2023, 97-99).

Gracias a esta distinción entre agencia moral y agencia funcional es posible superar la objeción de que castigar a las máquinas sería absurdo porque no pueden sufrir ni comprender el reproche. Tal objeción es válida contra quienes pretenden aplicar a la IA el mismo régimen de responsabilidad diseñado para seres humanos, pero no afecta a un modelo que reconoce desde el inicio la especificidad ontológica de los agentes artificiales y deriva de ella consecuencias normativas igualmente específicas. La sanción aplicada a un sistema de IA con agencia funcional no pretende infligir sufrimiento retributivo ni comunicar reproche moral, sino gestionar un riesgo mediante la modificación o eliminación de su fuente (Moch 2025, 298-300).

Ahora bien, el fundamento de la imputación a sistemas con agencia funcional no es la culpabilidad, categoría inaplicable a entidades sin conciencia, sino lo que puede denominarse «responsabilidad sistémica» o «imputabilidad técnica». Un sistema es imputable técnicamente cuando su operación satisface los criterios de autonomía, capacidad de ajuste normativo y desvinculación del control humano previamente señalados. Tal imputabilidad técnica fundamenta la aplicación de medidas preventivas sin necesidad de demostrar estados mentales inexistentes. La legitimidad de tales medidas deriva no del merecimiento del sistema, concepto sin aplicación aquí, sino de la necesidad de proteger bienes jurídicos frente a fuentes de riesgo que el ordenamiento no puede gestionar mediante otros instrumentos (La Vattiata 2023, 498-500).

Asimismo, la categoría de agencia penal funcional debe complementarse con mecanismos que impidan su utilización como escudo para eximir de responsabilidad a los actores humanos que efectivamente mantienen relaciones de control o beneficio respecto del sistema. El riesgo de que la personalidad algorítmica devenga en «cortafuegos de responsabilidad» para corporaciones y desarrolladores ha sido señalado reiteradamente como objeción contra cualquier forma de reconocimiento jurídico de la IA (Bryson *et al.* 2017, 280-288). Tal objeción resulta atendible y exige incorporar al modelo garantías que preserven la responsabilidad humana en los ámbitos donde efectivamente corresponde.

A modo de respuesta, la propuesta consiste en configurar la agencia funcional como complementaria y no sustitutiva de la responsabilidad humana. El reconocimiento del sistema como centro de imputación no excluye la responsabilidad concurrente de quienes participaron en su diseño, entrenamiento, implementación u operación, cuando dicha participación satisfaga independientemente los requisitos de la imputación tradicional. La estructura resultante es de responsabilidad distribuida: el sistema puede ser objeto de medidas preventivas en tanto fuente autónoma de riesgo, mientras que los actores humanos responden según los criterios ordinarios de culpabilidad cuando su conducta individual satisface los presupuestos del reproche (Teubner 2006, 515-520). Con ello, se evita tanto la impunidad derivada de la brecha de responsabilidad como la injusticia de sacrificar a chivos expiatorios humanos por fallos genuinamente sistémicos.

Una vez delineada la categoría de agencia penal funcional como instrumento teórico para integrar entidades no humanas en el campo de la imputación jurídica sin violentar los principios fundamentales del derecho penal, el paso siguiente consiste en traducir tal construcción conceptual en un diseño normativo operativo. En consecuencia, la sección siguiente presenta el modelo de imputación híbrida y responsabilidad distribuida, detallando su estructura regulatoria, los deberes correlativos de los actores involucrados y las medidas jurídicas aplicables a los sistemas algorítmicos con agencia funcional.

5. EL MODELO NORMATIVO DE IMPUTACIÓN HÍBRIDA Y RESPONSABILIDAD DISTRIBUIDA

La operacionalización de la agencia penal funcional requiere diseñar un marco normativo que articule los principios teóricos desarrollados en mecanismos institucionales concretos. En ese sentido, el modelo de imputación

híbrida y responsabilidad distribuida debe especificar los criterios de clasificación de sistemas según su nivel de autonomía, las obligaciones de los diversos actores del ciclo de vida algorítmico, las sanciones aplicables a cada tipo de entidad y los procedimientos para determinar la distribución de responsabilidades en casos concretos. El diseño debe garantizar simultáneamente la eficacia preventiva del sistema y su compatibilidad con los principios garantistas del derecho penal.

Con base en lo anterior, puede establecerse una taxonomía regulatoria de sistemas de IA según su relevancia para la imputación penal. La clasificación, inspirada en el enfoque basado en riesgos adoptado por el Reglamento Europeo de IA pero adaptada a las necesidades específicas del derecho penal, distingue cuatro categorías: sistemas instrumentales, sistemas asistidos, sistemas semiautónomos y sistemas autónomos de alto riesgo (Veale y Borgesius 2021, 100-108). Los sistemas instrumentales carecen de capacidad de aprendizaje y operan exclusivamente según instrucciones humanas explícitas; respecto de ellos aplican sin modificación las reglas tradicionales de responsabilidad por el producto. Los sistemas asistidos incorporan capacidades de aprendizaje pero operan bajo supervisión humana constante que retiene la decisión final; la responsabilidad recae en el supervisor salvo defectos de diseño imputables al fabricante. Los sistemas semiautónomos operan con autonomía en dominios delimitados pero mantienen mecanismos de intervención humana que pueden activarse ante situaciones anómalas; aquí comienza a ser relevante la agencia funcional del sistema como complemento de la responsabilidad humana residual. Los sistemas autónomos de alto riesgo satisfacen plenamente los criterios de agencia funcional y constituyen centros de imputación técnica diferenciados, sujetos al régimen de responsabilidad distribuida que a continuación se desarrolla.

En correspondencia con esta tipología, deben establecerse obligaciones diferenciadas para los actores del ciclo de vida del sistema algorítmico. Los desarrolladores asumen deberes de diligencia en el diseño, que incluyen la implementación de mecanismos de explicabilidad en la medida técnicamente posible, la realización de auditorías de seguridad y sesgo antes del despliegue, y la documentación suficiente de las capacidades y limitaciones del sistema (Weller 2021, 28-35). Los proveedores de datos de entrenamiento responden por la calidad, representatividad y licitud de los conjuntos de datos utilizados, debiendo implementar procedimientos de verificación que minimicen riesgos de sesgo discriminatorio o contenido ilícito.

Por otro lado, los implementadores, quienes adaptan sistemas genéricos a aplicaciones específicas, asumen responsabilidad por la adecuación del sistema al contexto de uso y por la configuración de parámetros operativos

dentro de rangos seguros. Los operadores responden por la supervisión continua del sistema durante su funcionamiento, la detección de anomalías y la activación de mecanismos de intervención cuando corresponda. Los usuarios finales, en tanto interactúan directamente con el sistema, asumen responsabilidad por el uso conforme a las instrucciones y limitaciones comunicadas (Yeung 2018, 512-518).

En línea con dicha distribución funcional, esta fragmentación de deberes refleja la condición estructural de la agencia en la era del Antropoceno, donde la capacidad de causar cambios en el mundo ya no reside en un sujeto soberano y unitario, sino que se encuentra diseminada en ensamblajes híbridos de humanos y no humanos (Arias Maldonado 2019, 68-74). En este contexto, la insistencia en mantener la responsabilidad humana no responde a una nostalgia humanista, sino a la necesidad política de evitar que la complejidad sistémica disuelva el vínculo entre decisión y consecuencia. El derecho penal debe asumir que la acción eficaz es hoy un producto colectivo y distribuido, lo que obliga a redefinir la autoría más allá de la ejecución material directa (Espinosa Rubio 2019, 25-30).

Ahora bien, la distribución de estas obligaciones no implica que todo incumplimiento genere automáticamente responsabilidad penal; este último paso requiere que la infracción del deber satisfaga adicionalmente los requisitos de imputación objetiva y subjetiva propios del tipo penal de que se trate. Lo que el modelo establece es un marco de deberes cuya infracción resulta relevante para el análisis de la responsabilidad cuando se producen resultados lesivos. La determinación de qué actor o actores responden en un caso concreto dependerá del análisis de las contribuciones causales, el cumplimiento o incumplimiento de los deberes específicos y la presencia de elementos subjetivos como dolo o imprudencia (Ibald 2024, 320-340).

En lo relativo a las sanciones aplicables a sistemas con agencia funcional, debe considerarse que estas sanciones, denominadas «medidas técnicas» para distinguirlas de las penas en sentido estricto, incluyen la desactivación temporal o permanente del sistema, la modificación obligatoria de parámetros o algoritmos, la restricción de ámbitos de operación, el reforzamiento de mecanismos de supervisión y la obligación de actualización de seguridad (Park 2024, 125-140). La graduación de estas medidas debe atender a la gravedad del daño causado o potencial, el grado de autonomía del sistema y la posibilidad técnica de corrección. La desactivación permanente, equivalente funcional de la incapacitación, queda reservada para sistemas cuya operación continuada representa riesgos intolerables que no pueden mitigarse mediante medidas menos invasivas.

En forma complementaria, las medidas técnicas pueden acompañarse de consecuencias patrimoniales que afectan no al sistema mismo sino a su titular o explotador. La constitución obligatoria de fondos de garantía o seguros de responsabilidad permite asegurar la compensación de las víctimas con independencia de la determinación de responsabilidades individuales, satisfaciendo la función reparadora sin perjuicio de las acciones de regreso que correspondan (Singh y Singh 2025, 8-15). La imposición de multas administrativas a las organizaciones que operan sistemas de alto riesgo sin cumplir las obligaciones de diligencia constituye un mecanismo de prevención general que incentiva la adopción de medidas de seguridad adecuadas.

En lo que respecta al procedimiento para determinar la distribución de responsabilidades cuando un sistema con agencia funcional causa un resultado lesivo, se requiere distinguir entre tres escenarios típicos: casos donde el daño deriva exclusivamente de decisiones autónomas del sistema que ningún actor humano podía anticipar ni evitar mediante el cumplimiento de sus deberes; casos donde el daño resulta del incumplimiento de deberes específicos por parte de uno o más actores humanos del ciclo de vida; luego casos mixtos donde concurren factores sistémicos y conductas humanas deficientes (Nikolaidis e Ionescu 2023, 50-51).

En el primer supuesto, la responsabilidad penal individual queda excluida por ausencia de base para el reproche, aplicándose exclusivamente las medidas técnicas respecto del sistema y, en su caso, las consecuencias patrimoniales derivadas de la responsabilidad objetiva del titular. Tal solución puede parecer insatisfactoria desde una perspectiva retributiva que exige identificar un culpable humano, pero resulta preferible a la alternativa de condenar injustamente a quienes no tenían control efectivo sobre el curso causal. En el segundo supuesto, la responsabilidad penal se imputa según las reglas ordinarias al actor cuya infracción del deber fundamenta el reproche, sin perjuicio de las medidas técnicas aplicables al sistema. En el tercer escenario, corresponde aplicar las reglas de concurrencia de causas y distribución de responsabilidades, pudiendo coexistir la sanción del actor humano con las medidas técnicas sobre el sistema y las consecuencias patrimoniales sobre el titular (Giannini 2023, 245-260).

Por lo que respecta a las reformas legislativas necesarias para implementar el modelo propuesto, estas deben operar en varios niveles: la modificación de la parte general del código penal para reconocer la categoría de agencia funcional y establecer las medidas técnicas como consecuencias jurídicas específicas; la creación de tipos penales específicos que sancionen las infracciones graves de los deberes del ciclo de vida algorítmico; la reforma procesal para establecer procedimientos adecuados de investigación de incidentes algorít-

micos, incluyendo el acceso a registros y la colaboración técnica obligatoria; la creación o fortalecimiento de autoridades administrativas especializadas con competencias de supervisión, certificación y sanción en el ámbito de la IA (Maskur *et al.* 2025, 1820-1835).

Con todo, la compatibilidad del modelo con los principios del derecho penal liberal exige atención específica. El principio de legalidad requiere que tanto la clasificación de sistemas como las medidas aplicables y los deberes de los actores del ciclo de vida estén definidos con suficiente precisión en normas previas al hecho. El principio de culpabilidad se preserva respecto de los actores humanos, cuya responsabilidad penal sigue exigiendo la concurrencia de elementos subjetivos; respecto de los sistemas con agencia funcional, este principio se sustituye por el criterio de imputabilidad técnica que fundamenta medidas preventivas sin contenido retributivo. El principio de proporcionalidad gobierna tanto la graduación de las medidas técnicas como la determinación de las sanciones a actores humanos, exigiendo adecuación entre la gravedad del injusto y la intensidad de la respuesta. El principio de subsidiariedad o *ultima ratio* orienta la decisión sobre cuándo resulta apropiada la intervención penal frente a mecanismos alternativos de control (Borgesano *et al.* 2025, 18-25).

Desde una perspectiva institucional, la implementación del modelo requiere finalmente una infraestructura de control *ex ante* y monitoreo continuo. Las agencias reguladoras especializadas deben estar dotadas de competencias técnicas para evaluar la conformidad de los sistemas con los requisitos de seguridad y transparencia, de capacidad sancionadora para imponer medidas correctivas ante incumplimientos y de recursos para realizar auditorías independientes de sistemas de alto riesgo. La experiencia de regímenes de seguridad industrial, farmacovigilancia o supervisión financiera ofrece modelos institucionales cuya adaptación al ámbito algorítmico resulta pertinente (Tong 2025, 16-20). Además, la cooperación internacional deviene esencial dado el carácter transfronterizo del desarrollo y despliegue de sistemas de IA, requiriendo armonización de estándares y mecanismos de reconocimiento mutuo entre jurisdicciones.

6. CONCLUSIONES

La disrupción que la IA autónoma introduce en el sistema penal no puede reducirse a una mera carencia normativa ni subsanarse mediante extensiones analógicas de categorías vigentes. Representa, más bien, la manifestación sintomática de un agotamiento estructural del paradigma antropocéntrico que

ha organizado históricamente los supuestos del reproche penal. El análisis genealógico y dogmático confirma que los pilares conceptuales de la teoría del delito, desde la voluntariedad aristotélica hasta la culpabilidad como juicio de reproche normativo, operan bajo una lógica de clausura ontológica que excluye, por principio, a cualquier entidad carente de experiencia consciente. No resulta viable, por tanto, insistir en la ficcionalización de un control humano que ya no existe en términos materiales ni proyectar categorías éticas sobre estructuras que operan por pura inferencia matemática. La alternativa exige abandonar la ontología del sujeto soberano y avanzar hacia un modelo relacional que reconozca la alteridad operativa de los sistemas algorítmicos como condición para su integración normativa.

Desde esta perspectiva, la elaboración de la agencia penal funcional constituye la contribución conceptual más relevante. Se trata de una categoría dogmática autónoma, desvinculada de la agencia moral y de la dignidad ontológica, que permite imputar consecuencias jurídicas a sistemas no humanos en virtud de su capacidad para actuar como fuentes independientes de riesgo sistémico. La síntesis propuesta, una imputación híbrida, evita tanto la antropomorfización ingenua como la impunidad estructural, habilitando un modelo de responsabilidad distribuida donde la llamada «personalidad» del algoritmo no responde a una sustancia metafísica, sino a una atribución funcional ajustada a criterios operativos de autonomía, opacidad y desvinculación del control humano.

A nivel normativo, esta conceptualización se traduce en un mandato de reconfiguración institucional orientado hacia una lógica de prevención y gestión de riesgos. Las sanciones técnicas, como la desactivación, modificación forzosa o restricción funcional, reemplazan la finalidad retributiva de la pena, operando como mecanismos de neutralización de peligros. La asignación de deberes positivos de diseño, auditoría y vigilancia a los actores humanos involucrados en el ciclo de vida del sistema transforma la opacidad técnica en una zona jurídicamente gobernada. La propuesta no permanece en el plano especulativo, sino que ofrece una arquitectura normativa concreta, orientada a reformas legislativas capaces de articular mecanismos de control *ex ante*, como la certificación algorítmica, con respuestas *ex post* que no requieran el imposible rastreo de intenciones en procesos estocásticos.

No obstante, la robustez teórica del modelo no agota los desafíos pendientes. Será imprescindible avanzar en la validación empírica de los umbrales de autonomía funcional propuestos, explorar la viabilidad técnica de los mecanismos de explicabilidad como condición previa a la imputación, y evaluar las tensiones constitucionales que puedan derivarse de la imposición de medidas coercitivas sobre activos digitales dotados de valor económico.

La dimensión global del fenómeno tecnológico añade un nivel adicional de complejidad, pues la eficacia de la agencia funcional como categoría jurídica dependerá, en última instancia, de su adopción coordinada en el plano internacional. La amenaza de paraísos regulatorios que erosionen la efectividad del modelo obliga a pensar en marcos de armonización normativa transfronteriza, así como en protocolos de cooperación institucional sostenidos en estándares técnicos compartidos.

El desafío epistemológico que plantea la IA invita a una relectura profunda de la filosofía del derecho penal. Abandonar el privilegio ontológico del sujeto humano no implica renunciar a los principios garantistas, sino asumir la madurez de un orden jurídico que debe operar en una realidad híbrida. La propuesta de una agencia funcional, lejos de ser un mero artificio técnico, representa un giro hacia una comprensión del derecho penal menos metafísica y más sistémica. Por consiguiente, la justicia del futuro dependerá de nuestra capacidad filosófica para disociar la imputación de la culpa moral, reconociendo que en la era del Antropoceno, marcada por la expansión tecnológica, el poder de actuar, así como la capacidad de causar daño, ya no constituye un atributo exclusivo de la conciencia humana.

REFERENCIAS BIBLIOGRÁFICAS

- AGUIRRE MARTÍNEZ, Guillermo. «Sobre los modelos encubiertos de antropocentrismo en nuestra relación con el medio orgánico». *Azafea. Revista de Filosofía* 21 (2019): 159-179.
- ARANCO BAGNASCO, Andrés. «Dignidad, constructivismo y fuentes de la normatividad: entre las experiencias del daño y las formas de menosprecio». *Azafea. Revista de Filosofía* 24 (2022): 165-186.
- ARIAS MALDONADO, Manuel. «La ética ecológica en el Antropoceno». *Azafea. Revista de Filosofía* 21 (2019): 55-76.
- ARISTÓTELES. *Ética a Nicómaco*. Trad. José Luis Calvo Martínez. Madrid: Alianza Editorial, 2001.
- ASIMOV, Isaac. «Runaround». *Astounding Science Fiction* 29, 1 (1942): 94-104.
- BASHIR MANGI, Danish; Irfan BUTRO y Tanveer AKHTAR MEMON. «AI and criminal liability: Theoretical dilemmas in applying criminal law to artificial intelligence». *The Critical Review of Social Sciences Studies* 3, 2 (2025): 2174-2186.
- BORGESANO, Francesco; Andrea DE MAIO; Pietro LAGHI y Roberto MUSMANNO. «Artificial intelligence and justice: A systematic literature review and future research perspectives on Justice 5.0». *European Journal of Innovation Management* (2025): publicación anticipada en línea.

- BRYSON, Joanna; Mihailis DIAMANTIS y Thomas GRANT. «Of, for, and by the people: The legal lacuna of synthetic persons». *Artificial Intelligence and Law* 25, 3 (2017): 273-291.
- BURRELL, Jenna. «How the machine ‘thinks’: Understanding opacity in machine learning algorithms». *Big Data & Society* 3, 1 (2016): 1-12.
- ÇINICI, Kaan. *KI als „Täter« oder Tatmittel: zur Strafbarkeit von Künstlicher Intelligenz*. Tesis de Diploma. Innsbruck: Universität Innsbruck, 2025.
- COECKELBERGH, Mark. *AI Ethics*. Cambridge, Mass.: MIT Press, 2020.
- CORRIDORI, Claudia. *Responsabilità penale per i reati commessi dall’Intelligenza Artificiale: «Machina delinquere non potest»*. Tesis de Maestría. Roma: LUISS Guido Carli, 2021.
- DESCARTES, René. *Meditaciones metafísicas*. Trad. Vidal Peña. Madrid: Alianza Editorial, 2010.
- DUFF, Robin Antony. *Answering for Crime: Responsibility and Liability in the Criminal Law*. Oxford: Hart Publishing, 2007.
- ELLAMEY, Youssef y Ahmed ELWAKAD. «The criminal responsibility of artificial intelligence systems: A prospective analytical study». *Corporate Law & Governance Review* 5, 1 (2023): 92-100.
- ESPINOSA RUBIO, Luciano. «Reflexiones sobre Antropoceno y colapso». *Azafea. Revista de Filosofía* 21 (2019): 11-31.
- FLORIDI, Luciano y Jeff W. SANDER. «On the morality of artificial agents». *Minds and Machines* 14, 3 (2004): 349-379.
- GIANNINI, Alessio. *Criminal behavior and accountability of artificial intelligence systems*. Tesis doctoral. Maastricht: Maastricht University / Florencia: Universidad de Florencia, 2023.
- GILABERT, Fernando. «Fundamentum scientiae: Heidegger y el no pensar de la ciencia moderna». *Azafea. Revista de Filosofía* 24 (2022): 187-199.
- HAMMOURI, Jihad Ahmad; Ahmad A. ALMAHASNEH; Khaled M. KHWAILEH y Mohammad M. AL-RAGGAD. «The criminal liability of artificial intelligence entities». *Pakistan Journal of Life and Social Sciences* 22, 2 (2024): 8785-8790.
- IBOLD, Victoria. «Strafrechtliche Produktverantwortung für KI-Produkte». En *Künstliche Intelligenz und Strafrecht: Zur strafrechtlichen Produktverantwortung in der Innovationsgesellschaft*, 219-422. Baden-Baden: Nomos, 2024.
- JAKOBS, Günther. *Strafrecht: Allgemeiner Teil*. 2.^a ed. Berlín: De Gruyter, 1991.
- KANT, Immanuel. *Crítica de la razón práctica*. Trad. Roberto R. Aramayo. Madrid: Alianza Editorial, 1998.
- LA VATTIATA, Franco Cataldo. «AI systems involved in harmful events: Liable persons or mere instruments? An interdisciplinary and comparative analysis». *Bio-Law Journal - Rivista di BioDiritto* 1 (2023): 485-502.
- LUHMANN, Niklas. *Sistemas sociales: Lineamientos para una teoría general*. Trad. Silvia Pappe y Brunhilde Erker. Barcelona: Anthropos / México: Universidad Iberoamericana / Bogotá: CEJA, 1998.

- LUHMANN, Niklas. *El derecho de la sociedad*. Trad. Javier Torres Nafarrate. México: Universidad Iberoamericana / Barcelona: Herder, 2005.
- MARX, Karl. *Manuscritos económicos y filosóficos de 1844*. Ed. Juan R. Fajardo. Biblioteca Virtual Espartaco, 2001.
- MASKUR, Muhammad Ali; Ali MASYHAR; Ratna DAMAYANTI; Dian Puspita RAMADA y Shubham SANYAL. «Reimagining criminal liability in the age of artificial intelligence: Toward a comparative and reform-oriented legal framework». *Journal of Law and Legal Reform* 6, 4 (2025): 1805-1838.
- MATTHIAS, Andreas. «The responsibility gap: Ascribing responsibility for the actions of learning automata». *Ethics and Information Technology* 6, 3 (2004): 175-183.
- MIR PUIG, Santiago. *Derecho Penal: Parte General*. 10.^a ed. Barcelona: Reppertor, 2016.
- MOCH, Eduard. «Autonomous systems (AI) and criminal imputability: Challenges for modern law». *International Journal of Research and Review* 12, 10 (2025): 291-301.
- NIEL, Luis. «Intencionalidad, objeto y sentido en la Gegenstandstheorie de Alexius Meinong». *Azafea. Revista de Filosofía* 17 (2015): 141-173.
- NIKOLAIDIS, Georgios y Andrei IONESCU. «The future of legal responsibility in human-AI collaboration: Revisiting fault, intent, and causation in hybrid decision structures». *Legal Studies in Digital Age* 2, 4 (2023): 48-60.
- PARK, Jiyeon. *Strafrechtliche Verantwortlichkeit im Zusammenhang mit autonomen Fahrzeugen: Problematik und Lösungsansätze*. Tesis doctoral. Berlín: Freie Universität Berlin, 2024.
- PROYAS, Alex, dir. *Yo, robot*. Estados Unidos: Twentieth Century Fox, 2004.
- ROMERO, Álvaro. «Dos concepciones de la identidad práctica: Aristóteles y Korsgaard». *Azafea. Revista de Filosofía* 27, 1 (2025): 287-310.
- ROXIN, Claus. *Derecho Penal: Parte General*. Tomo I, 2.^a ed. Madrid: Civitas, 1997.
- SCHERER, Matthew U. «Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies». *Harvard Journal of Law & Technology* 29, 2 (2016): 353-400.
- SIMONIGH, Chiara. «Anthropomorphism, cosmomorphism, metamorphosis: Between images and media environment». *Azafea. Revista de Filosofía* 19 (2017): 169-184.
- SINGH, Sanjay y Meera SINGH. «Criminal liability in AI-enabled autonomous vehicles: A comparative study». *SSRN Electronic Journal* (2025): s.p.
- TEUBNER, Gunther. «Rights of non-humans? Electronic agents and animals as new actors in politics and law». *Journal of Law and Society* 33, 4 (2006): 497-521.
- TONG, Mingxin. «The new actor: Artificial intelligence in criminology and criminal justice». *Criminology, Criminal Justice, Law & Society* 26, 2 (2025): 12-22.
- VEALE, Michael y Frederik Zuiderveen BORGESIU. «Demystifying the Draft EU Artificial Intelligence Act». *Computer Law Review International* 22, 4 (2021): 97-112.

- WALTERMANN, Antonia. «On the legal responsibility of artificially intelligent agents: Addressing three misconceptions». *Technology and Regulation* (2021): 28-38.
- WELLER, Adrian. «Transparency: Motivations and challenges». En *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, ed. Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen y Klaus-Robert Müller, 23-40. Cham: Springer, 2021.
- WELZEL, Hans. *Derecho Penal Alemán: Parte General*. 11.^a ed. Trads. Juan Bustos Ramírez y Sergio Yáñez Pérez. Santiago de Chile: Editorial Jurídica de Chile, 2002.
- YEUNG, Karen. «Algorithmic regulation: A critical interrogation». *Regulation & Governance* 12, 4 (2018): 505-523.