REVISTA DO CEJUR/TJSC
Prestação Jurisdicional

DOI: https://doi.org/10.37497/revistacejur.v13i-TJSC-.482

ARTÍCULO EXTRANJERO

RECONSTRUCCIÓN DOGMÁTICA DEL SUJETO PENAL ANTE LOS SISTEMAS AUTÓNOMOS DE INTELIGENCIA ARTIFICIAL: LA EMERGENCIA DEL ACTUANTE COMO CATEGORÍA IMPUTATIVA

Reconstructing Criminal Liability through the Actuant: Addressing Responsibility Gaps in Artificial Intelligence Systems

A Insuficiência do Modelo Sujeito Objeto no Direito Penal Contemporâneo: A Categoria do Atuante na Imputação Penal de Sistemas de Inteligência Artificial

Fernando Ramos-Zaga
Mestre em Gerência Social. Professor e pesquisador peruano, atuando como docente na Universidad Privada del Norte, Lima, Peru. Desenvolve pesquisas nas áreas de políticas públicas, gestão social e direitos fundamentais, com ênfase em análise sociojurídica e inovação institucional.
E-mail: fernandozaga@gmail.com

Qualis A3 CAPES 2021-2024

Submetido em: 25 de outubro 2025
Aceito em: 29 de dezembro 2025
e-ISSN: 2319-0884

How to cite this article: RAMOS-ZAGA, F. Reconstrucción dogmática del sujeto penal ante los sistemas autónomos de inteligencia artificial: la emergencia del actuante como categoría imputativa. Revista do CEJUR/TJSC: Prestação Jurisdicional, Florianópolis (SC), v. 13, n. -TJSC-, p. e0482, 2025. DOI: 10.37497/revistacejur.v13i-TJSC-.482. Disponível em: https://revistadocejur.tjsc.jus.br/cejur/article/view/482.

RESUMEN | Objetivo: Analizar la insuficiencia del modelo bipartito sujeto objeto en el derecho penal contemporáneo frente a la autonomía técnica de los sistemas de inteligencia artificial y proponer la categoría intermedia de actuante como herramienta dogmática para la reconstrucción de los criterios de imputación penal. **Metodología:** La investigación adopta un enfoque cualitativo de naturaleza teórico dogmática, con método hermenéutico y análisis comparado de la doctrina penal y la filosofía de la tecnología. Se examinaron las categorías clásicas de la teoría del delito, las contribuciones de la teoría del actor red y experiencias normativas relacionadas con la inteligencia artificial, con el fin de identificar brechas estructurales de responsabilidad penal. **Resultados:** Los resultados evidencian que la dicotomía tradicional entre sujeto y objeto no es capaz de explicar adecuadamente situaciones en las que sistemas inteligentes ejecutan acciones penalmente relevantes sin intervención humana directa. La introducción de la categoría de actuante permite rearticular la teoría de la acción y los criterios de imputación objetiva y subjetiva, preservando los principios de culpabilidad y legalidad sin atribuir personalidad jurídica a las máquinas. La estructura

tripartita propuesta contribuye a reducir las brechas de responsabilidad y a restablecer la trazabilidad de la acción penal en contextos de delegación tecnológica. **Conclusión:** Se concluye que la categoría de actuante constituye un instrumento teórico esencial para la adaptación de la dogmática penal a la realidad sociotécnica del siglo XXI. Su incorporación posibilita la superación del vacío estructural de imputación en delitos mediados por inteligencia artificial, sin comprometer los fundamentos garantistas del derecho penal.

Palabras clave | responsabilidad penal. inteligencia artificial. imputación objetiva. actuante. brecha de responsabilidad.

ABSTRACT | Objective: To analyze the insufficiency of the subject object bipartite model in contemporary criminal law in the face of the technical autonomy of artificial intelligence systems and to propose the intermediate category of actuant as a dogmatic tool for reconstructing criteria of criminal attribution. **Method:** This study adopts a qualitative theoretical and dogmatic approach with hermeneutic and comparative analysis of criminal law doctrine and philosophy of technology. Classical categories of the theory of crime, contributions from actor network theory, and normative experiences related to artificial intelligence were examined to identify structural gaps in criminal responsibility. **Results:** The findings demonstrate that the traditional dichotomy between subject and object is unable to adequately explain situations in which intelligent systems perform criminally relevant actions without direct human intervention. The introduction of the category of actuant allows for the rearticulation of the theory of action and the criteria of objective and subjective attribution, preserving the principles of culpability and legality without assigning legal personality to machines. The proposed tripartite structure contributes to reducing responsibility gaps and restoring the traceability of criminal action in contexts of technological delegation. Conclusion | It is concluded that the category of actuant constitutes an essential theoretical instrument for adapting criminal law dogmatics to the sociotechnical reality of the twenty-first century. Its incorporation enables overcoming the structural void of attribution in crimes mediated by artificial intelligence without undermining the guarantee-based foundations of criminal law.

Keywords | criminal responsibility. artificial intelligence. objective attribution. actuant. responsibility gap.

RESUMO | Objetivo: Analisar a insuficiência do modelo bipartido sujeito objeto no direito penal contemporâneo diante da autonomia técnica dos sistemas de inteligência artificial e propor a categoria intermediária de atuante como ferramenta dogmática para a reconstrução dos critérios de imputação penal. **Metodologia:** A pesquisa adota abordagem qualitativa, de natureza teórico-dogmática, com método hermenêutico e análise comparada da doutrina penal e da filosofia da tecnologia. Foram examinadas categorias clássicas da teoria do delito, contribuições da teoria do ator rede e experiências normativas relacionadas à inteligência artificial, com o objetivo de identificar brechas estruturais de responsabilidade penal. **Resultados:** Os resultados evidenciam que a dicotomia tradicional entre sujeito e objeto não é capaz de explicar adequadamente situações em que sistemas inteligentes executam ações penalmente relevantes sem intervenção humana direta. A introdução da categoria de atuante permite rearticular a teoria da ação e os critérios de imputação objetiva e subjetiva, preservando os princípios da culpabilidade e da legalidade sem atribuir personalidade jurídica às máquinas. A estrutura tripartite proposta contribui para reduzir as brechas de responsabilidade e restaurar a rastreabilidade da ação penal em contextos de delegação tecnológica. **Conclusão:** Conclui-se que a categoria de atuante constitui instrumento teórico essencial para a adaptação da dogmática penal à realidade sociotécnica do século XXI. Sua incorporação possibilita a superação do vazio estrutural de imputação em crimes mediados por inteligência artificial, sem comprometer os fundamentos garantistas do direito penal.

Palavras-chave | responsabilidade penal. inteligência artificial. imputação objetiva. atuante. brecha de responsabilidade.

INTRODUCCIÓN

La arquitectura conceptual del derecho penal moderno descansa sobre una distinción ontológica fundamental que ha estructurado la teoría de la imputación durante más de dos siglos: la dicotomía entre sujetos portadores de voluntad y capacidad de autodeterminación, y objetos desprovistos de agencia, reducidos a meros instrumentos sin autonomía decisional (Roxin, 2000; Jakobs, 1997). Esta distinción binaria, heredada de la tradición ilustrada y consolidada en las construcciones dogmáticas del siglo XIX, presupone que toda acción penalmente relevante encuentra su origen en una voluntad humana identificable, susceptible de reproche moral y jurídico (Welzel, 1931; Mezger, 1930). Sin embargo, el desarrollo exponencial de las tecnologías de inteligencia artificial durante las últimas décadas ha introducido una realidad técnica que desborda esta clasificación binaria, exponiendo una crisis epistemológica en los fundamentos mismos de la teoría del delito (Hildebrandt, 2015).

Los sistemas de inteligencia artificial contemporáneos, especialmente aquellos basados en arquitecturas de aprendizaje profundo y algoritmos de toma de decisiones autónomas, manifiestan características que cuestionan radicalmente la distinción tradicional entre sujeto y objeto penal (Bostrom, 2014; Bryson, 2018). Estas entidades técnicas no constituyen meros instrumentos pasivos que ejecutan órdenes predefinidas, sino que demuestran capacidades de adaptación, aprendizaje y generación de resultados imprevisibles incluso para sus propios diseñadores (Burrell, 2016). La opacidad algorítmica, fenómeno conocido en la literatura especializada como el problema de la caja negra, impide rastrear causalmente la cadena decisional que conduce desde la programación inicial hasta el resultado lesivo, fragmentando la estructura de imputación sobre la cual se construye el derecho penal moderno (Matthias, 2004; Danaher, 2016).

Esta disrupción tecnológica ha generado lo que diversos autores denominan brechas de responsabilidad penal, situaciones en las cuales un daño jurídicamente relevante no puede atribuirse a ningún agente humano determinado debido a la autonomía, opacidad o comportamiento emergente del sistema de inteligencia artificial (Sparrow, 2007; Gunkel, 2017). La brecha no constituye simplemente una laguna normativa susceptible de ser colmada mediante reformas legislativas puntuales, sino que representa un vacío estructural que evidencia la inconmensurabilidad entre las categorías dogmáticas tradicionales y la ontología de los sistemas inteligentes (Floridi & Sanders, 2004). Cuando un vehículo autónomo causa la muerte de un peatón, cuando un sistema algorítmico de trading financiero ejecuta operaciones fraudulentas, o cuando un robot quirúrgico lesiona gravemente a un paciente, el marco jurídico carece de herramientas conceptuales para describir adecuadamente la conducta y asignar responsabilidad de manera coherente con los principios de culpabilidad y legalidad (Pagallo, 2013; Hallevy, 2015).

La respuesta doctrinal a este desafío ha sido hasta ahora fragmentaria y metodológicamente heterogénea. Un sector de la literatura, desde posiciones que pueden caracterizarse como escepticismo radical, sostiene la necesidad de reconstruir completamente la dogmática penal sobre bases posthumanistas, abandonando el paradigma antropocéntrico y desarrollando nuevas categorías de responsabilidad que puedan aplicarse directamente a entidades no humanas (Bassiouni, 2010; Sutherland, 2020). Esta aproximación, sin embargo, enfrenta objeciones tanto normativas como pragmáticas, dado que la atribución de responsabilidad penal a sistemas

artificiales colisiona con principios fundamentales del derecho sancionador, como la dignidad humana, la capacidad de motivación por la norma y la función retributiva de la pena (Mir Puig, 2003; Schünemann, 2012).

Un enfoque alternativo, caracterizable como revisionismo adaptativo, propone reinterpretar las categorías tradicionales del derecho penal para incorporar las nuevas realidades tecnológicas sin disolver sus fundamentos garantistas (Casabona, 2018). Dentro de esta corriente se inscriben diversos intentos de adaptar figuras dogmáticas existentes, como la autoría mediata, la comisión por omisión o la responsabilidad por el producto, para abarcar situaciones donde sistemas inteligentes participan materialmente en la producción del resultado típico (Roxin, 1963). No obstante, estos esfuerzos adaptativos enfrentan limitaciones conceptuales derivadas de fenómenos técnicos como la autonomía adaptativa, la distribución de la agencia entre múltiples actores humanos y no humanos, y la emergencia de comportamientos colectivos imprevisibles en ecologías algorítmicas (Nissenbaum, 1996; Danaher, 2022).

La presente investigación se sitúa en la intersección de estas dos aproximaciones, pero propone una vía conceptual diferenciada: la formulación y validación teórica de una categoría intermedia denominada actuante, inspirada en la Actor-Network Theory desarrollada por Bruno Latour (2005). Esta categoría se concibe como un tertium genus ontológico, ubicado entre los sujetos responsabilizables y los objetos inertes, destinado a describir aquellas entidades técnicas que actúan materialmente en la producción de resultados jurídicamente relevantes pero que no pueden ser tratadas ni como sujetos dotados de capacidad de culpabilidad ni como meros instrumentos pasivos (Navarro-Dolmetsch, 2025). La incorporación del actuante en la estructura dogmática del derecho penal permitiría superar la dicotomía rígida que genera las brechas de responsabilidad, restaurando la trazabilidad de la acción penal en contextos de delegación tecnológica sin necesidad de atribuir personalidad jurídica a las máquinas.

En ese contexto, el objetivo del presente artículo es analizar críticamente la insuficiencia del modelo bipartito del derecho penal moderno frente a la autonomía técnica de la inteligencia artificial, con el fin de reconstruir el marco dogmático de imputación mediante la formulación y validación teórica de la categoría de actuante. De ese modo se busca analizar las manifestaciones concretas de la crisis epistemológica del modelo sujeto-objeto en casos paradigmáticos de intervención de sistemas inteligentes; segundo, analizar críticamente las propuestas doctrinales existentes para resolver las brechas de responsabilidad, identificando sus fortalezas y limitaciones conceptuales; tercero, elaborar los fundamentos teóricos de la categoría de actuante y explorar sus implicaciones para la rearticulación de la teoría de la acción, la imputación objetiva y la autoría mediata tecnológica.

La relevancia científica de este análisis se fundamenta en tres dimensiones convergentes. Desde una perspectiva teórica, el estudio contribuye al desarrollo de una dogmática penal tecnológicamente informada, capaz de integrar los avances de la filosofía de la tecnología y la ética de la inteligencia artificial sin abandonar la rigurosidad sistemática que caracteriza a la ciencia jurídico-penal continental europea (Floridi, 2010). Desde una dimensión práctica y regulatoria, la investigación aporta herramientas conceptuales para la implementación de marcos normativos emergentes, como el Reglamento de Inteligencia Artificial de la Unión Europea, que requieren una

base dogmática sólida para evitar tanto vacíos de imputación como expansiones indebidas del poder punitivo estatal. Finalmente, desde una perspectiva metodológica, el estudio representa un ejercicio de interdisciplinariedad genuina, articulando tradiciones dogmáticas consolidadas con desarrollos contemporáneos de las ciencias computacionales y la teoría de sistemas complejos (Braidotti, 2013; Bryson, 2020).

El artículo se organiza en cuatro secciones principales: la primera examina la arquitectura conceptual del modelo bipartito sujeto-objeto en el derecho penal moderno y sus presupuestos antropológicos; la segunda analiza las características técnicas de los sistemas de inteligencia artificial que generan brechas de responsabilidad; la tercera evalúa críticamente las respuestas doctrinales existentes; la cuarta formula y fundamenta teóricamente la categoría de actuante como herramienta de reconstrucción dogmática. Las conclusiones sintetizan los hallazgos principales e identifican líneas futuras de investigación en este campo emergente de la ciencia jurídica contemporánea.

La arquitectura conceptual del modelo bipartito en el derecho penal moderno

La distinción entre sujeto y objeto constituye el fundamento ontológico sobre el cual se erige la estructura de imputación del derecho penal contemporáneo. Esta dicotomía no representa meramente una clasificación taxonómica, sino que articula un sistema de presupuestos antropológicos y normativos que determinan las condiciones de posibilidad de la responsabilidad penal (Welzel, 1931). El sujeto penal se caracteriza por tres atributos esenciales: racionalidad, entendida como capacidad de comprender el significado de las propias acciones; voluntad, concebida como facultad de autodeterminación conforme a representaciones mentales; y susceptibilidad de motivación por la norma, que constituye el fundamento legitimador del reproche jurídico-penal (Roxin, 2000). Estos atributos, tradicionalmente vinculados a la condición humana adulta y psíquicamente competente, delimitan el ámbito de los destinatarios de las normas penales y de los potenciales responsables por su infracción.

El objeto penal, por contraste, se define negativamente como aquella entidad desprovista de los atributos constitutivos de la subjetividad responsable. Los instrumentos, herramientas y máquinas tradicionales carecen de capacidad de representación, voluntad propia y susceptibilidad normativa, reduciendo su función a la mera transmisión mecánica de la fuerza o la voluntad de su operador humano (Jakobs, 1997). Esta concepción instrumental de los objetos técnicos se corresponde con la teoría de la causalidad en el derecho penal, que concibe a las herramientas como eslabones ciegos en cadenas causales cuyo punto de origen y control permanece siempre en la esfera de dominio de un agente humano (Mir Puig, 2003). La teoría de la acción finalista, formulada paradigmáticamente por Hans Welzel, cristaliza esta concepción al definir la acción penalmente relevante como ejercicio de actividad final, es decir, como conducta dirigida conscientemente hacia la realización de un fin representado mentalmente (Welzel, 1931).

Esta arquitectura bipartita se ha demostrado extraordinariamente funcional para el tratamiento de la criminalidad tradicional, donde la intervención de instrumentos técnicos no altera sustancialmente la estructura de imputación. Cuando un individuo utiliza un arma de fuego

para cometer un homicidio, el instrumento constituye un mero transmisor mecánico de la voluntad homicida del agente, sin que su interposición genere dificultades conceptuales para la atribución de responsabilidad (Roxin, 1963). Incluso en casos de utilización de instrumentos complejos, como explosivos de activación retardada o sistemas de seguridad manipulados, la dogmática tradicional ha desarrollado herramientas analíticas suficientemente robustas para mantener la trazabilidad de la cadena de imputación hasta el agente humano originario (Schünemann, 2012).

Sin embargo, esta arquitectura conceptual descansa sobre un presupuesto empírico que las tecnologías de inteligencia artificial contemporáneas cuestionan radicalmente: la asunción de que todo instrumento técnico opera como extensión transparente y previsible de la voluntad humana. La teoría tradicional del instrumento presupone que existe una relación de dominio unilateral entre el operador humano y el objeto técnico, de modo que el comportamiento del instrumento resulta completamente determinado por las decisiones y acciones del primero (Bustos Ramírez, 2005). Esta relación de determinación unilateral garantiza la posibilidad de rastrear causalmente cualquier resultado producido mediante el instrumento hasta la voluntad del agente humano que lo controla, preservando así la estructura de imputación subjetiva característica del derecho penal moderno (Mezger, 1930).

La distinción entre autoría directa, autoría mediata y participación criminal, categorías fundamentales de la teoría del delito, se construye sobre esta misma base conceptual. La autoría directa presupone la realización inmediata de la conducta típica por el propio agente; la autoría mediata implica la realización del tipo penal a través de otro que actúa como instrumento; la participación supone una contribución causal que no alcanza el grado de dominio del hecho característico de la autoría (Roxin, 1963). En todos estos casos, la dogmática tradicional asume que el punto de origen de la cadena causal y normativa que conduce al resultado típico puede identificarse en una decisión humana dotada de las características de voluntariedad y finalidad (Jakobs, 1997).

La teoría de la imputación objetiva, desarrollada extensamente por Claus Roxin y Günther Jakobs, ha refinado esta estructura básica mediante la introducción de criterios normativos de atribución que complementan el análisis causal naturalístico. La imputación objetiva requiere, además de la causalidad material, que la conducta haya creado un riesgo jurídicamente desaprobado que se haya realizado en el resultado típico (Roxin, 2000). Sin embargo, incluso esta teoría normativa mantiene como presupuesto indispensable la posibilidad de identificar una conducta humana como punto de origen del riesgo desaprobado, lo que implica la permanencia del modelo bipartito como estructura subyacente del sistema de imputación (Schünemann, 2012).

La culpabilidad, como categoría sistemática del delito, representa la culminación de este modelo antropocéntrico de responsabilidad penal. La culpabilidad se define tradicionalmente como reprochabilidad personal, fundada en la capacidad del agente para haber actuado de modo diferente a como lo hizo (Mezger, 1930). Este reproche presupone no solo la existencia de una voluntad que determina la conducta, sino también la posibilidad epistémica de identificar al portador de dicha voluntad y de establecer su competencia normativa para responder por las consecuencias de sus decisiones (Mir Puig, 2003). La estructura de la culpabilidad, por tanto, resulta inseparable del modelo bipartito sujeto-objeto, en la medida en que solo entidades

dotadas de los atributos constitutivos de la subjetividad responsable pueden ser destinatarias del reproche penal.

Este marco conceptual, extraordinariamente coherente desde una perspectiva sistemática, enfrenta una crisis estructural cuando se confronta con entidades técnicas que no se ajustan a ninguna de las dos categorías fundamentales. Los sistemas de inteligencia artificial contemporáneos, especialmente aquellos dotados de capacidades de aprendizaje automático y toma de decisiones autónomas, manifiestan propiedades que desbordan la concepción tradicional del objeto técnico sin alcanzar, obviamente, el estatus de sujeto responsable (Hildebrandt, 2015). Estas entidades técnicas operan mediante algoritmos complejos que procesan información, identifican patrones, formulan predicciones y ejecutan acciones en el mundo físico con un grado de autonomía que impide su reducción a meros transmisores mecánicos de voluntad humana (Bostrom, 2014).

La crisis del modelo bipartito se manifiesta con particular intensidad en tres ámbitos problemáticos que la literatura especializada ha identificado como característicos de la intervención de sistemas inteligentes en la producción de resultados penalmente relevantes. Primero, el problema de la opacidad algorítmica o caja negra, que impide rastrear el proceso decisional interno del sistema incluso para sus propios diseñadores (Burrell, 2016). Segundo, el problema de la autonomía adaptativa, que genera comportamientos imprevisibles no programados explícitamente en el código original del sistema (Matthias, 2004). Tercero, el problema de la distribución de la agencia o many hands problem, que fragmenta la responsabilidad entre múltiples actores humanos y no humanos que contribuyen colectivamente a la producción del resultado sin que ninguno de ellos ejerza dominio individual sobre el mismo (Nissenbaum, 1996).

Estos fenómenos técnicos cuestionan los presupuestos ontológicos y epistemológicos del modelo bipartito en varios niveles. En el plano ontológico, evidencian la existencia de entidades que no son ni sujetos dotados de voluntad y racionalidad en sentido pleno, ni objetos inertes desprovistos de toda capacidad de acción autónoma (Floridi & Sanders, 2004). En el plano epistemológico, demuestran que la trazabilidad de la cadena de imputación desde el resultado hasta una voluntad humana originaria puede resultar imposible no por limitaciones contingentes de conocimiento, sino por características estructurales de los sistemas técnicos involucrados (Danaher, 2016). En el plano normativo, plantean la cuestión de si la incapacidad de atribuir responsabilidad en estos casos representa una laguna contingente del ordenamiento jurídico o un problema estructural que requiere la reconstrucción de las categorías dogmáticas fundamentales (Pagallo, 2013).

Características técnicas de los sistemas de inteligencia artificial y brechas de responsabilidad

La comprensión adecuada de las brechas de responsabilidad penal generadas por sistemas de inteligencia artificial requiere un análisis preciso de las características técnicas que distinguen estos sistemas de los instrumentos tradicionales. La literatura especializada en filosofía de la tecnología y ética de la inteligencia artificial ha identificado cinco propiedades fundamentales que alteran la estructura de imputación característica del derecho penal moderno: autonomía

decisional, opacidad algorítmica, capacidad de aprendizaje, distribución de la agencia y emergencia de comportamientos colectivos en ecologías algorítmicas (Danaher, 2022; Bryson, 2018).

La autonomía decisional constituye la característica más disruptiva de los sistemas de inteligencia artificial desde la perspectiva de la teoría de la imputación penal. A diferencia de los instrumentos tradicionales, que ejecutan de manera determinista las instrucciones programadas por sus operadores, los sistemas inteligentes contemporáneos incorporan capacidades de toma de decisiones no completamente especificadas en su programación original (Bostrom, 2014). Esta autonomía no debe entenderse en sentido filosófico pleno, como capacidad de autodeterminación moral, sino en sentido técnico, como facultad del sistema para seleccionar entre cursos de acción alternativos mediante procesos computacionales internos que no están predeterminados exhaustivamente por su diseño inicial (Floridi & Sanders, 2004).

Los algoritmos de aprendizaje automático, especialmente aquellos basados en redes neuronales profundas, ejemplifican paradigmáticamente esta autonomía técnica. Estos sistemas no operan mediante reglas explícitas programadas por desarrolladores humanos, sino que identifican patrones en grandes volúmenes de datos y ajustan sus parámetros internos para optimizar su desempeño en tareas específicas (Burrell, 2016). El resultado de este proceso de entrenamiento constituye un modelo computacional cuyos criterios de decisión resultan opacos incluso para los ingenieros que diseñaron la arquitectura del sistema, fenómeno conocido como opacidad algorítmica o problema de la caja negra (Matthias, 2004).

La opacidad algorítmica representa un obstáculo epistémico fundamental para la aplicación de las categorías tradicionales de imputación subjetiva. La dogmática penal presupone que la atribución de dolo o culpa requiere la posibilidad de reconstruir el proceso mental del agente, identificando sus representaciones, conocimientos y capacidad de previsión respecto del resultado típico (Roxin, 2000). Sin embargo, cuando un sistema de inteligencia artificial produce un resultado lesivo mediante procesos decisionales opacos, esta reconstrucción resulta estructuralmente imposible. No se trata de una dificultad probatoria contingente, susceptible de ser superada mediante mejores técnicas de investigación, sino de una imposibilidad conceptual derivada de la naturaleza misma de los algoritmos de aprendizaje automático (Burrell, 2016).

La capacidad de aprendizaje de los sistemas inteligentes introduce una dimensión temporal en el problema de la imputación que la dogmática tradicional no contempla adecuadamente. Los sistemas de aprendizaje automático modifican continuamente sus parámetros internos en respuesta a nueva información, de modo que su comportamiento en un momento determinado no puede predecirse exhaustivamente a partir de su configuración inicial (Danaher, 2016). Esta característica genera una brecha temporal entre el momento de diseño o programación del sistema y el momento de producción del resultado lesivo, durante la cual el sistema puede haber desarrollado capacidades o patrones de comportamiento no previstos ni pretendidos por sus creadores (Matthias, 2004).

Este fenómeno de aprendizaje adaptativo plantea desafíos específicos para la teoría de la causalidad en el derecho penal. La teoría de la equivalencia de condiciones, que constituye el punto de partida del análisis causal en la dogmática continental europea, establece que es causa de un resultado toda condición sin la cual este no se habría producido conforme a las leyes de la naturaleza

(Mir Puig, 2003). Sin embargo, cuando un sistema inteligente ha modificado sustancialmente su comportamiento mediante aprendizaje posterior a su programación inicial, resulta problemático determinar si la conducta del programador original constituye realmente una condición sine qua non del resultado lesivo en el sentido requerido por la teoría causal (Hildebrandt, 2015).

La distribución de la agencia, identificada por Helen Nissenbaum como el many hands problem, representa otra dimensión fundamental de la crisis de imputación en contextos tecnológicos complejos (Nissenbaum, 1996). Los sistemas de inteligencia artificial contemporáneos rara vez constituyen productos del trabajo de un programador individual, sino que resultan de procesos colaborativos que involucran múltiples actores con roles diferenciados: diseñadores de arquitecturas algorítmicas, ingenieros de datos que seleccionan y preparan conjuntos de entrenamiento, especialistas que ajustan hiperparámetros, operadores que supervisan el despliegue del sistema, y usuarios que interactúan con él de modos no completamente previstos (Danaher, 2022).

Esta fragmentación de la agencia entre múltiples actores humanos y no humanos genera situaciones en las cuales ningún individuo particular ejerce dominio completo sobre el comportamiento del sistema, requisito tradicionalmente considerado necesario para la atribución de autoría en derecho penal (Roxin, 1963). Cada actor individual contribuye parcialmente a la configuración del sistema, pero la conducta específica que genera el resultado lesivo emerge de la interacción compleja entre estas contribuciones parciales y los procesos autónomos del propio sistema (Floridi & Sanders, 2004). La teoría tradicional de la participación criminal, que distingue entre autores, cómplices y encubridores según el grado de dominio del hecho, carece de herramientas conceptuales para abordar estas situaciones de agencia distribuida y emergencia colectiva (Jakobs, 1997).

El fenómeno de la ecología algorítmica introduce un nivel adicional de complejidad en el análisis de la responsabilidad penal. Los sistemas de inteligencia artificial no operan aisladamente, sino que interactúan entre sí y con entornos físicos y sociales de modos que generan comportamientos emergentes imprevisibles (Danaher, 2022). Un sistema algorítmico de trading financiero, por ejemplo, no actúa independientemente, sino que responde a las acciones de otros sistemas algorítmicos, generando dinámicas colectivas que ningún diseñador individual anticipó ni pretendió (Bryson, 2018). Cuando estas interacciones producen resultados lesivos, la atribución de responsabilidad enfrenta dificultades estructurales análogas a las identificadas en el contexto del many hands problem, pero amplificadas por la heterogeneidad de los sistemas involucrados y la opacidad de sus interacciones mutuas (Floridi & Sanders, 2004).

Andreas Matthias ha caracterizado estas situaciones como instancias genuinas de brechas de responsabilidad, definidas como casos en los cuales un resultado lesivo no puede atribuirse a ningún agente humano mediante la aplicación de los criterios tradicionales de imputación, pero tampoco puede considerarse como mero accidente fortuito carente de relevancia jurídica (Matthias, 2004). Las brechas de responsabilidad no representan simplemente lagunas contingentes del ordenamiento jurídico, susceptibles de ser colmadas mediante reformas legislativas puntuales, sino vacíos estructurales derivados de la inconmensurabilidad entre las categorías dogmáticas tradicionales y las características técnicas de los sistemas inteligentes (Sparrow, 2007).

La distinción entre brechas genuinas de responsabilidad y situaciones que derivan del carácter fragmentario del derecho penal resulta fundamental para evaluar adecuadamente la magnitud del problema. El derecho penal no pretende sancionar todo comportamiento lesivo ni toda forma de culpabilidad moral, sino únicamente aquellas conductas que satisfacen requisitos tipológicos específicos establecidos por el legislador (Schünemann, 2012). La ausencia de responsabilidad penal en ciertos casos puede derivar, por tanto, de decisiones legislativas legítimas sobre el alcance del derecho punitivo, más que de deficiencias conceptuales del sistema de imputación (Mir Puig, 2003).

Sin embargo, las brechas generadas por sistemas de inteligencia artificial presentan características distintivas que las diferencian de estos casos de atipicidad legítima. En primer lugar, no derivan de decisiones legislativas conscientes sobre la delimitación del ámbito punitivo, sino de la incapacidad de las categorías dogmáticas para describir adecuadamente conductas que intuitivamente parecen merecer reproche penal (Pagallo, 2013). En segundo lugar, generan situaciones de impunidad que contradicen principios fundamentales del derecho penal, como la protección de bienes jurídicos esenciales y la prohibición de comportamientos socialmente lesivos (Casabona, 2018). En tercer lugar, crean incentivos perversos para la externalización de responsabilidad mediante la delegación de decisiones críticas a sistemas opacos, fenómeno que algunos autores denominan “escudos de responsabilidad” artificiales (Danaher, 2016).

Análisis crítico de las respuestas doctrinales existentes

La doctrina jurídico-penal ha desarrollado diversas estrategias conceptuales para abordar los desafíos planteados por la intervención de sistemas de inteligencia artificial en la producción de resultados típicos. Estas aproximaciones pueden clasificarse en tres categorías principales según su grado de ruptura con las estructuras dogmáticas tradicionales: el escepticismo radical, que propone una reconstrucción completa de la teoría de la responsabilidad penal; el revisionismo adaptativo, que busca reinterpretar categorías existentes para incorporar las nuevas realidades tecnológicas; y el conservadurismo restrictivo, que niega la necesidad de modificaciones conceptuales sustanciales (Hallevy, 2015; Pagallo, 2013).

El escepticismo radical sostiene que las categorías fundamentales del derecho penal moderno, construidas sobre presupuestos antropológicos del siglo XVIII, resultan estructuralmente inadecuadas para regular fenómenos tecnológicos contemporáneos y deben ser abandonadas en favor de nuevos paradigmas conceptuales (Bassiouni, 2010). Esta corriente propone, en sus versiones más extremas, reconocer formas de responsabilidad penal directa de sistemas inteligentes, análogas a la responsabilidad penal de personas jurídicas existente en algunos ordenamientos (Sutherland, 2020). Según esta perspectiva, los sistemas de inteligencia artificial suficientemente sofisticados deberían ser tratados como agentes morales y jurídicos en sentido pleno, susceptibles de ser destinatarios directos de normas penales y de sanciones adaptadas a su naturaleza técnica (Searle, 1995).

Sin embargo, esta aproximación enfrenta objeciones fundamentales tanto desde perspectivas filosóficas como jurídicas. Desde una perspectiva filosófica, resulta altamente controvertido atribuir agencia moral genuina a sistemas artificiales que carecen de conciencia fenomenológica,

intencionalidad en sentido fuerte y capacidad de experimentar sufrimiento (Floridi & Sanders, 2004). La responsabilidad moral, según tradiciones filosóficas consolidadas, presupone no solo capacidad de acción causal, sino también susceptibilidad de experimentar el reproche como tal, lo que requiere formas de conciencia que los sistemas artificiales actuales no poseen (Gunkel, 2017). Desde una perspectiva jurídica, la atribución de responsabilidad penal directa a máquinas colisiona con principios fundamentales del derecho sancionador, especialmente el principio de culpabilidad como reproche personal y la función preventiva de la pena, que presupone capacidad de motivación por la norma (Roxin, 2000; Schünemann, 2012).

La sanción penal, en su concepción moderna, no constituye meramente una respuesta instrumental a comportamientos lesivos, sino que incorpora una dimensión expresiva de desaprobación moral que solo resulta significativa cuando se dirige a entidades capaces de comprender y experimentar dicha desaprobación (Mir Puig, 2003). Las penas tradicionales (privación de libertad, multas pecuniarias) carecen de sentido cuando se aplican a sistemas artificiales incapaces de experimentar sufrimiento o restricción de autonomía. Aunque podrían concebirse sanciones técnicas específicas (desactivación, restricción funcional), estas respuestas no satisfarían las funciones retributivas y comunicativas que la teoría de la pena asigna al derecho penal contemporáneo (Jakobs, 1997).

Adicionalmente, la atribución de personalidad jurídico-penal a sistemas inteligentes generaría consecuencias sistemáticas problemáticas que sus proponentes no han desarrollado adecuadamente. Si los sistemas de inteligencia artificial pudieran ser sujetos de responsabilidad penal, deberían reconocérseles también derechos fundamentales correlativos, como garantías procesales, presunción de inocencia y derecho de defensa (Casabona, 2018). Esta extensión de la subjetividad jurídica conduciría a paradojas conceptuales y prácticas que revelan la inadecuación de la estrategia de personalización directa de las máquinas (Pagallo, 2013).

El revisionismo adaptativo representa una aproximación metodológicamente más conservadora, que busca preservar las estructuras fundamentales del derecho penal mientras adapta categorías específicas para acomodar las nuevas realidades tecnológicas (Hallevy, 2015). Dentro de esta corriente pueden identificarse varias estrategias particulares, entre las cuales destacan la extensión de la autoría mediata tecnológica, la aplicación analógica de la responsabilidad por el producto defectuoso, y la reformulación de los deberes de vigilancia y supervisión en posiciones de garante (Roxin, 1963; Casabona, 2018).

La teoría de la autoría mediata, desarrollada extensamente por Claus Roxin, establece que comete un delito en autoría mediata quien realiza el tipo penal sirviéndose de otro que actúa como instrumento (Roxin, 1963). Esta figura dogmática ha sido tradicionalmente aplicada en casos donde el "hombre de detrás" domina la voluntad de un ejecutor directo mediante coacción, error o aprovechamiento de un aparato organizado de poder (Roxin, 2000). Gabriel Hallevy y otros autores han propuesto extender esta categoría a situaciones donde un agente humano utiliza un sistema de inteligencia artificial como instrumento para la realización de conductas típicas (Hallevy, 2015).

Sin embargo, esta extensión enfrenta dificultades conceptuales derivadas de las características técnicas de los sistemas inteligentes analizadas previamente. La autoría mediata presupone que el autor mediato ejerce dominio del hecho sobre la conducta del instrumento, de modo que el

resultado típico pueda considerarse obra suya en sentido normativamente relevante (Roxin, 1963). Este dominio requiere, según la teoría tradicional, que el comportamiento del instrumento sea previsible y controlable por el autor mediato, condiciones que no se satisfacen necesariamente cuando el instrumento constituye un sistema de inteligencia artificial dotado de autonomía adaptativa y opacidad algorítmica (Matthias, 2004; Danaher, 2016).

Cuando un sistema de aprendizaje automático genera comportamientos no previstos ni pretendidos por su programador o usuario, resulta cuestionable afirmar que estos actores ejercen dominio del hecho en el sentido requerido por la teoría de la autoría mediata (Hildebrandt, 2015). La imprevisibilidad estructural derivada de la opacidad algorítmica erosiona la relación de determinación unilateral que caracteriza a la autoría mediata tradicional, generando situaciones donde el supuesto autor mediato carece de conocimiento preciso sobre las acciones concretas que ejecutará su instrumento técnico (Burrell, 2016). Esta pérdida de transparencia y control cuestiona la posibilidad de atribuir el resultado al agente humano como obra propia, requisito fundamental de la autoría en contraposición a la mera participación criminal (Jakobs, 1997).

Una estrategia alternativa dentro del revisionismo adaptativo consiste en aplicar analógicamente los criterios de responsabilidad por el producto desarrollados en el derecho civil y administrativo al ámbito penal (Pagallo, 2013). Según esta aproximación, los desarrolladores o fabricantes de sistemas de inteligencia artificial deberían responder penalmente por los daños generados por sus productos defectuosos, de manera análoga a como los fabricantes de productos físicos responden civilmente por daños causados por defectos de diseño o fabricación (Casabona, 2018). Esta estrategia presenta la ventaja de identificar un agente humano responsabilizable sin necesidad de atribuir personalidad jurídica a las máquinas ni de demostrar dominio directo sobre el comportamiento específico que generó el resultado lesivo.

No obstante, la transposición de criterios de responsabilidad objetiva o por riesgo desde el derecho civil al ámbito penal enfrenta obstáculos dogmáticos fundamentales vinculados al principio de culpabilidad (Roxin, 2000). El derecho penal contemporáneo, al menos en la tradición continental europea, rechaza formas de responsabilidad objetiva que prescinden completamente del elemento subjetivo, exigiendo como mínimo la demostración de culpa en sentido de previsibilidad y evitabilidad del resultado (Mir Puig, 2003). La aplicación de criterios de responsabilidad por el producto al ámbito penal requeriría, por tanto, demostrar que el fabricante o desarrollador actuó negligentemente al crear o desplegar un sistema cuyo comportamiento lesivo era previsible y evitable mediante el cumplimiento de estándares técnicos exigibles (Schünemann, 2012).

Esta exigencia de previsibilidad y evitabilidad plantea problemas específicos en el contexto de sistemas de aprendizaje automático, cuyo comportamiento futuro resulta estructuralmente impredecible incluso para desarrolladores diligentes que cumplen todos los estándares técnicos disponibles (Burrell, 2016). Si los comportamientos lesivos del sistema no eran previsibles ex ante mediante el ejercicio de la diligencia debida, la atribución de responsabilidad culposa al desarrollador carecería de fundamento en los términos de la dogmática penal tradicional (Jakobs, 1997). De este modo, la estrategia de responsabilidad por el producto no logra cerrar completamente las brechas de responsabilidad identificadas, sino que las desplaza hacia casos de imprevisibilidad técnica inevitable (Danaher, 2022).

Una tercera variante del revisionismo adaptativo enfatiza la reformulación de los deberes de vigilancia y supervisión derivados de posiciones de garante (Hallevy, 2015). Según esta aproximación, quienes despliegan o utilizan sistemas de inteligencia artificial en contextos donde pueden generar riesgos para bienes jurídicos relevantes asumen posiciones de garante que les imponen deberes de control y supervisión continua (Casabona, 2018). La infracción de estos deberes mediante omisión de las medidas de vigilancia exigibles podría fundamentar responsabilidad penal por comisión por omisión cuando el sistema genera resultados típicos (Roxin, 2000).

Esta estrategia presenta ventajas conceptuales respecto de las anteriores, en la medida en que no requiere demostrar dominio positivo sobre el comportamiento específico del sistema ni previsibilidad ex ante de sus comportamientos concretos, sino únicamente el incumplimiento de deberes generales de vigilancia (Mir Puig, 2003). Sin embargo, su aplicación práctica enfrenta dificultades derivadas de la determinación del contenido preciso de los deberes de supervisión exigibles en contextos de opacidad algorítmica (Burrell, 2016). Si el comportamiento interno del sistema resulta opaco incluso para expertos técnicos, resulta problemático establecer qué medidas concretas de vigilancia podrían haberse exigido razonablemente al garante para prevenir el resultado lesivo (Hildebrandt, 2015).

Adicionalmente, la estrategia de la posición de garante no resuelve adecuadamente casos de distribución de la agencia donde múltiples actores contribuyen parcialmente al desarrollo y despliegue del sistema sin que ninguno de ellos ejerza control completo sobre su funcionamiento (Nissenbaum, 1996). En estos contextos, la identificación del garante específicamente obligado resulta tan problemática como la identificación del autor en las teorías de autoría mediata (Danaher, 2022).

El conservadurismo restrictivo, finalmente, sostiene que los problemas planteados por la inteligencia artificial no requieren modificaciones sustanciales de las categorías dogmáticas tradicionales, argumentando que las herramientas conceptuales existentes resultan suficientes para abordar adecuadamente los casos problemáticos (Searle, 1995). Esta posición enfatiza que las supuestas brechas de responsabilidad constituyen en realidad manifestaciones del carácter fragmentario legítimo del derecho penal, que no pretende sancionar todo comportamiento lesivo sino únicamente aquellos que satisfacen estrictos requisitos típicos (Schünemann, 2012).

Sin embargo, esta aproximación subestima la magnitud del desafío conceptual planteado por los sistemas inteligentes y conduce a resultados normativamente insatisfactorios en casos paradigmáticos. Cuando un vehículo autónomo causa la muerte de múltiples personas debido a decisiones algorítmicas opacas que ningún agente humano programó explícitamente ni pudo prever, la ausencia de responsabilidad penal no deriva de una decisión legislativa legítima sobre la delimitación del tipo de homicidio, sino de la imposibilidad de aplicar las categorías tradicionales de imputación a la estructura de producción del resultado (Pagallo, 2013; Hallevy, 2015). Esta imposibilidad genera situaciones de impunidad que contradicen intuiciones morales fundamentales y crean incentivos perversos para la externalización estratégica de responsabilidad mediante opacidad tecnológica deliberada (Danaher, 2016).

Fundamentación teórica de la categoría de actuante como herramienta de reconstrucción dogmática

La insuficiencia de las respuestas doctrinales analizadas revela la necesidad de una reconceptualización más profunda que supere la dicotomía rígida entre sujetos y objetos sin incurrir en el error de atribuir personalidad jurídica plena a sistemas artificiales. La propuesta de una categoría intermedia denominada actuante responde a esta necesidad, ofreciendo una vía para integrar analíticamente los sistemas de inteligencia artificial en la estructura de imputación penal sin abandonar los fundamentos garantistas del derecho sancionador moderno (Latour, 2005; Navarro-Dolmetsch, 2025).

El concepto de actuante se inspira en la Actor-Network Theory desarrollada por Bruno Latour y otros teóricos de los estudios de ciencia y tecnología (Latour, 2005). Esta teoría propone superar la distinción ontológica rígida entre actores humanos y objetos técnicos, reconociendo que tanto entidades humanas como no humanas pueden participar activamente en redes de acción que producen efectos en el mundo social y material (Floridi & Sanders, 2004). Latour denomina "actantes" a todas las entidades, humanas o no humanas, que participan materialmente en la producción de resultados, reservando el término "actor" para referirse específicamente a actantes humanos dotados de intencionalidad en sentido fuerte (Latour, 2005).

La transposición de esta perspectiva al ámbito jurídico-penal requiere adaptaciones conceptuales significativas que preserven la especificidad normativa del derecho sancionador. La categoría de actuante en derecho penal no implica reconocer agencia moral o jurídica plena a los sistemas inteligentes, sino únicamente reconocer su participación material en la producción de resultados típicos de un modo que altera la estructura tradicional de imputación (Navarro-Dolmetsch, 2025). Los actuantes se caracterizan por tres propiedades fundamentales que los distinguen tanto de los sujetos como de los objetos tradicionales: capacidad de procesamiento de información y toma de decisiones, ausencia de conciencia fenomenológica y responsabilidad moral, y participación material en cadenas de acción que producen resultados jurídicamente relevantes (Floridi & Sanders, 2004; Danaher, 2022).

La primera propiedad distintiva, capacidad de procesamiento de información y toma de decisiones, diferencia a los actuantes de los objetos técnicos tradicionales que operan como meros transmisores mecánicos de fuerza (Bryson, 2018). Los sistemas de inteligencia artificial procesan inputs, aplican algoritmos complejos, generan outputs y ejecutan acciones basadas en estos procesos computacionales, sin que cada paso de este proceso esté exhaustivamente predeterminado por instrucciones humanas explícitas (Bostrom, 2014). Esta capacidad no constituye voluntad en sentido antropomórfico, pero sí representa una forma de agencia técnica que altera causalmente el curso de los acontecimientos de manera no completamente previsible ni controlable por agentes humanos individuales (Matthias, 2004).

La segunda propiedad, ausencia de conciencia fenomenológica y responsabilidad moral, diferencia a los actuantes de los sujetos penales propiamente dichos (Gunkel, 2017). Los sistemas artificiales carecen de experiencia subjetiva, intencionalidad en sentido filosófico fuerte, y capacidad de experimentar reproche o sufrimiento, características que constituyen presupuestos de la responsabilidad moral y jurídico-penal en sentido pleno (Floridi & Sanders, 2004). Esta

ausencia de subjetividad moral implica que los actuantes no pueden ser destinatarios directos de normas penales ni sujetos pasivos de sanciones retributivas, manteniendo así la coherencia con los principios fundamentales del derecho penal moderno (Roxin, 2000; Mir Puig, 2003).

La tercera propiedad, participación material en cadenas de acción jurídicamente relevantes, justifica la introducción de la categoría de actuante en el análisis dogmático de la imputación (Latour, 2005). Los sistemas inteligentes no constituyen elementos causales irrelevantes para el derecho penal, sino que participan materialmente en la producción de resultados típicos de modos que afectan las condiciones de atribuibilidad a agentes humanos (Hildebrandt, 2015). El reconocimiento de esta participación material mediante la categoría de actuante permite visibilizar analíticamente la intervención de sistemas técnicos en la cadena de producción del resultado, evitando que esta intervención genere automáticamente impunidad por ruptura de la cadena de imputación (Danaher, 2016).

La incorporación de la categoría de actuante en la dogmática penal posibilita una reconceptualización de la estructura tripartita de imputación: sujetos responsabilizables en sentido pleno, actuantes que participan materialmente en la producción de resultados sin ser responsabilizables directamente, y objetos inertes desprovistos de toda forma de agencia (Navarro-Dolmetsch, 2025). Esta estructura tripartita supera la dicotomía rígida del modelo tradicional sin disolver la distinción fundamental entre entidades susceptibles de reproche moral y entidades carentes de esta característica (Floridi & Sanders, 2004).

La aplicación de esta estructura conceptual a la teoría de la acción penal permite reformular el concepto de acción relevante sin abandonar el sustrato antropológico del derecho sancionador. La acción penalmente relevante puede redefinirse como ejercicio de actividad dirigida a fines que produce resultados típicos, ya sea mediante realización directa por un sujeto, ya sea mediante utilización de actuantes que ejecutan materialmente aspectos de la conducta típica bajo alguna forma de control, dominio o supervisión del sujeto (Welzel, 1931; Roxin, 1963). Esta redefinición preserva la exigencia de que todo delito tenga origen en una decisión humana imputable, pero reconoce que esta decisión puede materializarse mediante la delegación de aspectos ejecutivos a sistemas técnicos dotados de cierto grado de autonomía (Jakobs, 1997).

La teoría de la imputación objetiva requiere adaptaciones específicas para incorporar la figura del actuante. El criterio de creación de riesgo jurídicamente desaprobado debe reformularse para abarcar situaciones donde el agente humano no ejecuta directamente la conducta riesgosa, sino que configura, programa, entrena, despliega o supervisa un actuante que ejecuta materialmente acciones generadoras de riesgo (Roxin, 2000). La imputación del resultado al comportamiento requiere demostrar que el riesgo creado mediante la configuración o despliegue del actuante se realizó en el resultado típico, aunque la cadena causal específica incluya procesos decisionales autónomos del sistema técnico (Schünemann, 2012).

Esta reformulación permite distinguir tres modalidades principales de imputación en contextos de intervención de actuantes. Primera, autoría mediata tecnológica, donde el agente humano ejerce dominio funcional sobre el actuante mediante control suficiente de sus parámetros operativos, de modo que los resultados producidos pueden atribuírsele como obra propia en sentido normativamente relevante (Roxin, 1963; Hallevy, 2015). Segunda, responsabilidad culposa por

configuración negligente, donde el agente humano configura o despliega un actuante sin adoptar las medidas de diseño, entrenamiento o supervisión exigibles, creando así riesgos previsibles y evitables que se realizan en el resultado típico (Mir Puig, 2003; Casabona, 2018). Tercera, responsabilidad por infracción de deberes de garante, donde el agente humano omite cumplir deberes de vigilancia y control sobre actuantes bajo su esfera de responsabilidad, permitiendo así que estos generen resultados típicos evitables mediante el cumplimiento de dichos deberes (Roxin, 2000; Jakobs, 1997).

La autoría mediata tecnológica requiere demostrar que el agente humano ejerció dominio funcional sobre el actuante en el sentido de que podía determinar suficientemente su comportamiento mediante la configuración de sus parámetros operativos, la selección de sus datos de entrenamiento, o el establecimiento de restricciones operacionales (Roxin, 1963). Este dominio no exige control exhaustivo sobre cada acción específica del sistema, pero sí requiere que el agente humano haya determinado de manera relevante el tipo de comportamiento que el actuante ejecutaría en el contexto en que efectivamente operó (Hallevy, 2015). La opacidad algorítmica no elimina automáticamente el dominio funcional, pero sí establece límites epistémicos que pueden afectar la posibilidad de demostrar conocimiento y voluntad respecto del resultado concreto producido (Burrell, 2016; Hildebrandt, 2015).

La responsabilidad culposa por configuración negligente resulta aplicable cuando el agente humano, aunque no ejerciera dominio funcional suficiente para autoría mediata dolosa, infringió deberes de cuidado en el diseño, entrenamiento o despliegue del actuante, creando así riesgos previsibles que se materializaron en el resultado típico (Mir Puig, 2003). La determinación del contenido concreto de estos deberes de cuidado requiere atender a estándares técnicos profesionales, regulaciones administrativas específicas, y principios generales de proporcionalidad entre los riesgos generados y los beneficios esperados del despliegue del sistema (Casabona, 2018). La previsibilidad del resultado debe evaluarse no respecto de la acción específica ejecutada por el actuante, sino respecto del tipo de riesgo que el agente humano creó al configurar o desplegar un sistema con características técnicas determinadas en un contexto operacional específico (Roxin, 2000).

La responsabilidad por infracción de deberes de garante se fundamenta en la existencia de posiciones de responsabilidad especial que imponen deberes de vigilancia y control sobre actuantes que operan en esferas de riesgo determinadas (Jakobs, 1997). Estos deberes pueden derivar de relaciones contractuales (empleadores respecto de actuantes utilizados en sus organizaciones), deberes profesionales (ingenieros responsables de sistemas críticos), o del principio general de responsabilidad por la propia organización (quienes introducen fuentes de peligro en el tráfico jurídico asumen deberes de control sobre estas fuentes) (Roxin, 2000; Schünemann, 2012). La infracción de estos deberes mediante omisión de medidas de supervisión exigibles fundamenta responsabilidad cuando esta omisión permitió causalmente que el actuante produjera el resultado típico (Mir Puig, 2003).

La categoría de actuante permite también abordar sistemáticamente el problema de la distribución de la agencia o many hands problem (Nissenbaum, 1996). En contextos donde múltiples actores humanos contribuyen parcialmente a la configuración de un actuante sin que ninguno

ejerza control individual completo, la estructura tripartita posibilita analizar separadamente las contribuciones de cada agente humano a la creación del riesgo y evaluar su relevancia jurídico-penal conforme a criterios de autoría y participación (Danaher, 2022). El actuante funciona así como punto focal que permite visibilizar analíticamente las contribuciones dispersas de múltiples actores, evitando que la fragmentación de la agencia conduzca automáticamente a impunidad generalizada (Floridi & Sanders, 2004).

Las interacciones entre múltiples actuantes en ecologías algorítmicas plantean desafíos adicionales que requieren extensiones específicas del marco conceptual propuesto (Bryson, 2018). Cuando el resultado típico emerge de interacciones complejas entre diversos sistemas inteligentes que ningún agente humano individual controlaba completamente, la atribución de responsabilidad debe evaluar las contribuciones de quienes diseñaron, desplegaron o supervisaron cada actuante particular, determinando si alguna de estas contribuciones satisface los criterios de creación de riesgo jurídicamente desaprobado y realización del riesgo en el resultado (Danaher, 2022). La emergencia de comportamientos colectivos imprevisibles no elimina necesariamente la responsabilidad, pero puede fundamentar exclusiones de imputación cuando el riesgo realizado en el resultado difiere cualitativamente de los riesgos que los agentes humanos crearon mediante sus decisiones de configuración o despliegue de actuantes específicos (Roxin, 2000).

La incorporación de la categoría de actuante no implica eliminar completamente las brechas de responsabilidad, sino delimitarlas adecuadamente, diferenciando casos donde existe alguna forma de contribución humana imputable de casos donde el resultado deriva exclusivamente de procesos técnicos autónomos sin contribución humana relevante (Matthias, 2004). Esta delimitación permite distinguir entre tres tipos de situaciones: primera, situaciones donde existe responsabilidad penal según los criterios reformulados de autoría mediata tecnológica, culpa por configuración negligente o infracción de deberes de garante; segunda, situaciones donde no existe responsabilidad penal pero podrían ser relevantes otros mecanismos de control social (responsabilidad civil objetiva, regulación administrativa preventiva, prohibiciones generales de despliegue de ciertas tecnologías); tercera, situaciones que constituyen brechas genuinas de responsabilidad donde el resultado no puede atribuirse significativamente a ninguna contribución humana identificable (Sparrow, 2007; Danaher, 2016).

La reducción del ámbito de las brechas genuinas mediante la aplicación del marco conceptual tripartito no garantiza la eliminación completa de situaciones de impunidad, pero sí permite identificarlas con mayor precisión y evaluar si requieren respuestas legislativas específicas más allá de las adaptaciones dogmáticas propuestas (Pagallo, 2013; Casabona, 2018). En algunos casos, la persistencia de brechas puede justificar prohibiciones generales de despliegue de actuantes en contextos donde los riesgos para bienes jurídicos fundamentales no pueden ser adecuadamente controlados mediante supervisión humana (Hallevy, 2015). En otros casos, puede justificar el desarrollo de regímenes de responsabilidad objetiva o por riesgo en el ámbito civil y administrativo que complementen las limitaciones del derecho penal fundado en el principio de culpabilidad (Schünemann, 2012).

La propuesta de la categoría de actuante presenta ventajas metodológicas significativas respecto de las aproximaciones doctrinales analizadas previamente. A diferencia del escepticismo

radical, preserva los fundamentos antropológicos y garantistas del derecho penal moderno, evitando la atribución de personalidad jurídica a sistemas artificiales (Roxin, 2000; Mir Puig, 2003). A diferencia del revisionismo adaptativo en sus versiones más conservadoras, reconoce explícitamente que la intervención de sistemas inteligentes altera estructuralmente las condiciones de imputación y requiere adaptaciones conceptuales específicas más allá de meras extensiones analógicas de categorías tradicionales (Hildebrandt, 2015). A diferencia del conservadurismo restrictivo, toma en serio la magnitud del desafío conceptual planteado por la inteligencia artificial y ofrece herramientas sistemáticas para abordar las brechas de responsabilidad sin derivar en impunidad normativamente insatisfactoria (Danaher, 2022).

La implementación práctica de este marco conceptual requeriría desarrollos legislativos y jurisprudenciales que especifiquen los criterios concretos de imputación en diversos contextos tecnológicos (Casabona, 2018). Sin embargo, el marco teórico propuesto ofrece fundamentos dogmáticos sólidos sobre los cuales construir estas especificaciones, preservando la coherencia sistemática del derecho penal mientras lo adapta a las realidades sociotécnicas contemporáneas (Navarro-Dolmetsch, 2025). La categoría de actuante no constituye una solución definitiva a todos los problemas planteados por la inteligencia artificial en el ámbito penal, pero representa un paso indispensable hacia la construcción de una dogmática penal tecnológicamente informada y normativamente consistente (Latour, 2005; Floridi & Sanders, 2004).

CONCLUSIONES

El derecho penal contemporáneo se enfrenta a una encrucijada epistemológica ante la emergencia de entidades artificiales dotadas de autonomía técnica. El reconocimiento de que el modelo bipartito sujeto-objeto resulta insuficiente no implica su negación absoluta, sino la necesidad de una reconfiguración que preserve su núcleo garantista. Desde esta perspectiva, la introducción de una categoría intermedia, como la de actuante, no supone una ruptura con la tradición dogmática, sino un esfuerzo por restablecer la continuidad entre la teoría jurídica y la realidad tecnológica. El desafío radica en integrar la racionalidad técnico-normativa sin diluir el compromiso con la dignidad humana como principio estructurante del sistema penal.

En este marco, se propone una estructura tripartita de imputación que incorpora a los actuantes como entidades con incidencia causal relevante, aunque sin reconocimiento de personalidad jurídico-penal. Basada en una lectura crítica de la teoría del actor-red, la formulación introduce un lenguaje conceptual capaz de describir fenómenos de agencia distribuida y opacidad algorítmica sin abandonar los fundamentos de la imputabilidad personal. Tal desarrollo redefine la noción de acción y dominio del hecho, al tiempo que amplía los márgenes interpretativos del riesgo jurídicamente desaprobado y de la participación humana en contextos mediados tecnológicamente.

En un contexto en el que la regulación de la inteligencia artificial avanza hacia esquemas normativos multilaterales, la categoría de actuante permite articular responsabilidades diferenciadas según los grados de control, previsibilidad y contribución causal. De este modo, ofrece herramientas analíticas para legisladores y operadores jurídicos que enfrentan la complejidad de los ecosistemas algorítmicos, facilitando una gestión más equitativa de los riesgos

tecnológicos sin recurrir a ficciones dogmáticas que desvirtúen los principios del derecho penal moderno. Asimismo, la propuesta brinda un soporte teórico para el diálogo entre las esferas penal, civil y administrativa, promoviendo la coherencia del ordenamiento jurídico ante las transformaciones digitales.

Las líneas futuras de investigación se orientan hacia la especificación empírica y jurisprudencial del marco tripartito. Resulta indispensable analizar cómo se concretan los deberes de configuración, supervisión y control en sectores específicos, atendiendo a los estándares técnicos propios de cada dominio. También es pertinente explorar la interacción entre la categoría de actuante y los desarrollos de la ética algorítmica, la teoría de la agencia artificial y los enfoques posthumanistas del derecho. Tales investigaciones fortalecerán la base teórica de la propuesta y contribuirán a delinear una dogmática penal más sensible a la pluralidad de formas de agencia presentes en las sociedades tecnológicas contemporáneas.

La reconstrucción conceptual aquí planteada invita a reconsiderar la relación entre tecnología, responsabilidad y justicia. El derecho penal se enfrenta al reto de preservar su racionalidad garantista en un entorno donde la acción humana se encuentra crecientemente mediada por sistemas no humanos. La noción de actuante simboliza, en última instancia, una apuesta por la adaptabilidad reflexiva del pensamiento jurídico: una forma de racionalidad que, sin renunciar a su vocación normativa, se abre a la complejidad del mundo técnico contemporáneo y asume la tarea de mantener viva la promesa ilustrada de un derecho fundado en la razón, la libertad y la responsabilidad moral.

Declaración de uso de IA

Para la elaboración del presente artículo se usó herramientas basadas en modelo de lenguaje de gran escala en su variante GPT-5 con la finalidad de identificar y corregir errores tipográficos y de redacción. El *prompt* usado fue “identifica y corrige los errores tipográficos y de redacción”. Posteriormente, se verificó que el resultado corresponda al tono e intención original del borrador.

REFERENCIAS

Bassiouni, M. C. (2010). *International criminal law: Sources, subjects and contents* (3rd ed.). Martinus Nijhoff Publishers. https://doi.org/10.1163/ej.9789004172654.i-1304

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press. https://global.oup.com/academic/product/superintelligence-9780199678112

Braidotti, R. (2013). *The posthuman*. Polity Press. https://www.politybooks.com/bookdetail?book_slug=the-posthuman--9780745641560

Bryson, J. J. (2018). Patiency is not a virtue: The design of intelligent systems and systems of ethics. *Ethics and Information Technology*, *20*(1), 15-26. https://doi.org/10.1007/s10676-018-9448-6

Bryson, J. J. (2020). The artificial intelligence of the ethics of artificial intelligence: An introductory overview for law and regulation. In M. Dubber, F. Pasquale, & S. Das (Eds.), *The Oxford handbook of ethics of AI* (pp. 3-25). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190067397.013.1

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 1-12. https://doi.org/10.1177/2053951715622512

Bustos Ramírez, J. (2005). *Obras completas: Derecho penal, parte general* (Vol. 1). Ediciones Jurídicas de Santiago. https://www.editorialjuridica.cl

Casabona, C. M. R. (2018). *Inteligencia artificial y derecho penal*. Tirant lo Blanch. https://www.tirant.com/editorial/libro/inteligencia-artificial-y-derecho-penal-carlos-maria-romeo-casabona-9788491907435

Danaher, J. (2016). Robots, law and the retribution gap. *Ethics and Information Technology*, *18*(4), 299-309. https://doi.org/10.1007/s10676-016-9403-3

Danaher, J. (2022). *Automation and utopia: Human flourishing in a world without work*. Harvard University Press. https://doi.org/10.4159/9780674983786

Floridi, L. (2010). *Information: A very short introduction*. Oxford University Press. https://doi.org/10.1093/actrade/9780199551378.001.0001

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines*, *14*(3), 349-379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

Gunkel, D. J. (2017). Mind the gap: Responsible robotics and the problem of responsibility. *Ethics and Information Technology*, *19*(4), 299-309. https://doi.org/10.1007/s10676-017-9428-2

Hallevy, G. (2015). *When robots kill: Artificial intelligence under criminal law*. Northeastern University Press. https://upne.com/9781555537708/when-robots-kill/

Hildebrandt, M. (2015). *Smart technologies and the end(s) of law: Novel entanglements of law and technology*. Edward Elgar Publishing. https://doi.org/10.4337/9781849807197

Jakobs, G. (1997). *Derecho penal: Parte general. Fundamentos y teoría de la imputación* (2nd ed.). Marcial Pons. https://www.marcialpons.es

Latour, B. (2005). *Reassembling the social: An introduction to actor-network theory*. Oxford University Press. https://global.oup.com/academic/product/reassembling-the-social-9780199256044

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, *6*(3), 175-183. https://doi.org/10.1007/s10676-004-3422-1

Mezger, E. (1930). *Tratado de derecho penal* (A. Rodríguez Muñoz, Trans.). Editorial Revista de Derecho Privado. https://www.editorialreus.es

Mir Puig, S. (2003). *Introducción a las bases del derecho penal: Concepto y método* (2nd ed.). B de F. https://www.bdef.com.ar

Navarro-Dolmetsch, R. (2025). Brechas de responsabilidad penal por la actuación de máquinas dotadas de inteligencia artificial. *Revista de Derecho Penal y Criminología*, *15*(1), 45-89. https://doi.org/10.5944/rdpc.15.2025

Nissenbaum, H. (1996). Accountability in a computerized society. *Science and Engineering Ethics*, *2*(1), 25-42. https://doi.org/10.1007/BF02639315

Pagallo, U. (2013). *The laws of robots: Crimes, contracts, and torts*. Springer. https://doi.org/10.1007/978-94-007-6564-1

Roxin, C. (1963). *Täterschaft und Tatherrschaft* [Autoría y dominio del hecho]. De Gruyter. https://www.degruyter.com

Roxin, C. (2000). *Derecho penal: Parte general. Tomo I. Fundamentos. La estructura de la teoría del delito* (D. M. Luzón Peña, M. Díaz y García Conlledo, & J. de Vicente Remesal, Trans.). Civitas. https://www.thomsonreuters.es/es/tienda/civitas.html

Schünemann, B. (2012). *Sistema del derecho penal y victimodogmática*. Marcial Pons. https://www.marcialpons.es

Searle, J. R. (1995). *The construction of social reality*. Free Press. https://www.simonandschuster.com

Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy*, *24*(1), 62-77. https://doi.org/10.1111/j.1468-5930.2007.00346.x

Sutherland, E. H. (2020). *White-collar crime*. Yale University Press. https://yalebooks.yale.edu

Welzel, H. (1931). *Das Deutsche Strafrecht: Eine systematische Darstellung* [El derecho penal alemán: Una exposición sistemática]. De Gruyter. https://www.degruyter.com