ISSN: 0101-9562
ISSN ELETRÔNICO: 2177-7055

SEQÜÊNCIA

Estudos jurídicos e políticos

Publicação do
Programa de Pós-Graduação em Direito da UFSC

VOLUME 47 ▪ ANO 2026

UFSC PPGD

SciELO

CAPES QUALIS A1

SEQÜÊNCIA – ESTUDOS JURÍDICOS E POLÍTICOS é uma publicação temática e de periodicidade quadrimestral, editada pelo Programa de Pós-Graduação Stricto Sensu em Direito da Universidade Federal de Santa Catarina – UFSC.

SEQÜÊNCIA – ESTUDOS JURÍDICOS E POLÍTICOS is a thematic publication, printed every four months, edited by the Program in law of the Federal University of Santa Catarina – UFSC.

Versão eletrônica: http://www.periodicos.ufsc.br/index.php/sequencia

Editora-Chefe: Norma Sueli Padilha

Editor Associado: José Sérgio da Silva Cristóvam

Editores Adjuntos: Priscilla Camargo Santos, Thanderson Pereira de Sousa

A publicação é indexada nas seguintes bases de dados e diretórios/
The Publication is indexed in the following databases and directories:

Base OJS
Base PKP
CCN (Catálogo Coletivo Nacional)
Dialnet
DOAJ (Directory of Open Access Journals)
EBSCOhost
Genamics Journalseek
Google Scholar
ICAP (Indexação Compartilhada de Artigos de Periódicos)
Latindex
LivRe!
ÖAW
OJS
PKP
Portal de Periódicos UFSC
Portal do SEER
ProQuest
SciELO
Scopus/Elsevier
Sherpa/Romeo
Sumarios.org
ULRICH'S
vLex

Ficha catalográfica

Seqüência: Estudos jurídicos e políticos. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Direito. n.1 (janeiro 1980)-.

Florianópolis: Fundação José Boiteux. 1980-.

Publicação contínua

Resumo em português e inglês

Versão impressa ISSN 0101-9562

Versão on-line ISSN 2177-7055

1. Ciência jurídica. 2. Teoria política. 3. Filosoia do direito. 4. Periódicos. I. Universidade Federal de Santa Catarina. Programa de Pós-graduação em Direito

CDU 34(05)

Catalogação na fonte por: João Oscar do Espírito Santo CRB 14/849

SEQÜÊNCIA

Publicação do
Programa de Pós-Graduação em Direito da UFSC

Estudos jurídicos
e políticos

Volume 47

e109148
https://doi.org/10.5007/2177-7055.2026.e109148

ARTÍCULO ORIGINAL

De la autonomía emergente a la personalidad instrumental: reconstrucción dogmática de la responsabilidad en agentes inteligentes de frontera

Emergent autonomy and the collapse of causal traceability: reconstituting functional liability for frontier model-based agents

Fernando A. Ramos-Zaga[1]
[1]Universidad Privada del Norte, Lima, Perú.

Resumen: Este artículo analiza las tensiones estructurales que la opacidad epistémica inherente a los modelos de frontera empleados en agentes inteligentes produce sobre las garantías constitucionales de motivación, trazabilidad y control jurisdiccional. La investigación parte de la constatación de que la autonomía emergente e imposibilidad técnica de reconstruir cadenas causales en la operatividad de estos sistemas hacen inviable el paradigma del control humano significativo, convirtiéndolo en una ficción normativa incapaz de asegurar una responsabilidad efectiva. A partir de este diagnóstico, se adopta un enfoque teórico-dogmático de orientación crítico-analítica para comprender la crisis de legitimidad provocada por la inteligencia artificial generativa. Los resultados muestran que la ausencia de sujetos humanos identificables vinculados causalmente a los efectos de estos modelos genera un vacío estructural de imputación, incompatible con las exigencias del Estado de Derecho. Se concluye que la creación de una personalidad jurídica algorítmica de carácter instrumental, inspirada en antecedentes como el *peculium* romano, constituye una alternativa normativamente coherente para restablecer centros funcionales de responsabilidad, garantizar la reparación y preservar la coherencia del sistema jurídico en contextos de agencia distribuida.

Palabras clave: Personalidad jurídica algorítmica. Agencia funcional. Agentes inteligentes. Control humano significativo.

Abstract: This article examines the structural tensions that the epistemic opacity inherent in frontier models deployed in AI agents imposes on the constitutional

Direito autoral e licença de uso: Este artigo está sob uma Licença Creative Commons. Com essa licença, você pode compartilhar, adaptar, para qualquer fim, desde que atribua a autoria da obra e forneça um link para a licença, e indicar se foram feitas alterações.

guarantees of reason-giving, traceability, and judicial oversight. The research begins with the premise that emergent autonomy and the technical infeasibility of reconstructing causal chains in the operation of these systems render the paradigm of meaningful human control structurally unviable, reducing it to a normative fiction incapable of securing effective accountability. From this diagnosis, the study adopts a critical-analytical and legal-dogmatic approach to address the legitimacy crisis triggered by generative artificial intelligence. The findings demonstrate that the absence of identifiable human agents causally linked to the effects produced by these models creates a structural void of legal imputation, incompatible with the normative demands of the Rule of Law. The article concludes that the establishment of an instrumental form of AI legal personhood constitutes a normatively coherent alternative to reconstitute functional centers of liability, ensure effective redress, and preserve the internal coherence of the legal system in contexts shaped by distributed agency.

Keywords: AI legal personhood. Functional agency. AI agents. Meaningful human control.

1 INTRODUCCIÓN

La introducción de agentes inteligentes (en adelante, AGI) implementados sobre modelos de lenguaje de frontera en procesos institucionales que afectan derechos fundamentales plantea un desafío estructural para los sistemas jurídicos contemporáneos. A diferencia de los algoritmos tradicionales, cuyo funcionamiento podía ser auditado mediante trazabilidad lógica y causal, los sistemas actuales operan mediante inferencias probabilísticas en arquitecturas distribuidas cuya dinámica interna no resulta accesible ni siquiera para sus desarrolladores. La opacidad epistémica, lejos de ser un defecto subsanable, constituye una propiedad inherente de la tecnología y se manifiesta en la imposibilidad de reconstruir cadenas causales, prever comportamientos emergentes o asignar responsabilidad conforme a esquemas tradicionales de imputación. La consecuencia inmediata de dicha transformación no es solo técnica, sino también política y jurídica: el fundamento epistémico que sostiene la legitimidad de las decisiones públicas se ve gravemente erosionado.

Desde una perspectiva teórica, el problema se sitúa en la intersección entre teoría constitucional, dogmática de la responsabilidad y estudios sobre Inteligencia Artificial (en adelante, IA). La literatura ha explorado los riesgos que plantea la automatización de decisiones en contextos como la salud, la justicia o la administración pública (Citron, 2008; Char et al., 2018), así como la insuficiencia de los mecanismos tradicionales de imputación frente a sistemas con autonomía emergente (Teubner, 2018). En paralelo, los marcos regulatorios han intentado preservar la agencia humana mediante el principio de "control humano significativo" (Santoni de Sio & van den Hoven, 2018), aunque estudios recientes muestran su inviabilidad estructural ante la escala y opacidad de los modelos actuales (Veluwenkamp, 2022; Elish, 2019). Sin embargo, tales abordajes suelen fragmentarse entre aproximaciones técnico-descriptivas y postulados normativos abstractos, sin articular una respuesta dogmática coherente que permita al derecho adaptarse a una realidad de agencia algorítmica distribuida.

Dentro de este panorama, persiste una brecha relevante en la literatura jurídica: no se ha sistematizado adecuadamente cómo la opacidad estructural de los modelos de frontera impide no solo la supervisión humana efectiva, sino también el cumplimiento de garantías constitucionales mínimas como la motivación de las decisiones, la contradicción procesal o el control jurisdiccional. Tampoco se ha desarrollado una alternativa dogmática que permita imputar consecuencias jurídicas a entidades que, sin ser sujetos de derecho en sentido clásico, generan efectos materiales y simbólicos en la esfera pública. Tal ausencia conceptual limita gravemente la capacidad del derecho para responder normativamente a daños reales provocados por sistemas que actúan fuera de las estructuras tradicionales de control.

La investigación se justifica, por tanto, en la necesidad urgente de dotar al derecho de herramientas teóricas que le permitan mantener su función ordenadora en entornos de creciente complejidad técnica. La estrategia de atribuir responsabilidad a operadores humanos que no controlan efectivamente el comportamiento de los sistemas, además

de jurídicamente insostenible, genera zonas de deformación moral donde se asigna culpa sin poder real de intervención (Elish, 2019). En lugar de insistir en modelos ficticios de supervisión o extender indiscriminadamente la responsabilidad objetiva, se requiere una reformulación dogmática que reoriente la arquitectura de imputación jurídica hacia centros funcionales que reflejen la nueva distribución de la agencia.

La hipótesis que se explora plantea que la atribución de una personalidad jurídica algorítmica limitada puede constituir un mecanismo normativo viable para restaurar la imputación efectiva en contextos donde la acción humana directa se ha diluido fácticamente. Tal figura permitiría vincular a los modelos autónomos con patrimonios de afectación específicos, sin necesidad de reconocerles derechos subjetivos ni exonerar a los humanos implicados en su desarrollo. El objetivo es puramente instrumental: garantizar solvencia para el resarcimiento, delimitar responsabilidades funcionales y preservar la coherencia interna del sistema jurídico. No se propone una ruptura teórica con la tradición jurídica, sino una adaptación mediante la extensión controlada de categorías ya existentes, como el *peculium* romano (Teubner, 2018) o las figuras de entes sin personalidad plena pero con capacidad patrimonial.

A la luz del escenario contemporáneo, donde las decisiones algorítmicas comienzan a configurar el acceso a bienes públicos, la tutela de derechos fundamentales y el contenido mismo de las normas jurídicas, la respuesta institucional no puede limitarse a ajustes procedimentales o exhortaciones éticas. La aceleración tecnológica impone al derecho el deber de revisar críticamente sus supuestos antropocéntricos, sin abdicar de sus funciones de garantía y control. La pregunta no es si debe reconocerse personalidad a los sistemas, sino cómo debe estructurarse la atribución de consecuencias jurídicas en contextos donde ya producen efectos. En tal sentido, la investigación se sitúa en el centro de una discusión que interpela tanto al constitucionalismo como a la teoría general del derecho.

En consecuencia, el objetivo de este estudio es analizar las tensiones estructurales que la opacidad epistémica inherente a los modelos de frontera empleados en agentes inteligentes produce sobre las garantías constitucionales de motivación, trazabilidad y control jurisdiccional, con el propósito de fundamentar dogmáticamente la viabilidad de una personalidad jurídica algorítmica como mecanismo jurídico funcional. La contribución reside en ofrecer un marco conceptual coherente que permita al derecho recuperar su capacidad de imputación en entornos de alta complejidad técnica, evitando tanto la negación simplista de la autonomía algorítmica como la atribución ficticia de responsabilidad humana en ausencia de condiciones epistémicas para ejercerla.

2 FUNDAMENTOS TÉCNICOS QUE IMPOSIBILITAN LA TRAZABILIDAD CAUSAL

La comprensión rigurosa de las tensiones que el uso de AGI implementados sobre modelos de lenguaje de frontera generan en el ordenamiento jurídico exige, como paso previo ineludible, una caracterización técnica de su arquitectura computacional. Tal aproximación no constituye un ejercicio de divulgación científica accesorio, sino un presupuesto metodológico fundamental, las propiedades que definen el comportamiento de dichos sistemas, autonomía emergente y opacidad epistémica, no son contingencias técnicas susceptibles de corrección ingenieril, sino características ontológicas intrínsecas derivadas de su diseño arquitectónico. La ruptura de la linealidad causal que producen opera en un plano fáctico anterior a cualquier valoración normativa, condicionando estructuralmente las posibilidades del derecho para observarlos, regularlos e imputarles consecuencias.

En continuidad con ello, los modelos de lenguaje de frontera (*frontier LLMs*), pertenecen a la familia de sistemas de IA generativa basados en arquitecturas de *transformers* cuyo funcionamiento se sustenta en mecanismos de autoatención distribuida que procesan secuencias

de tokens mediante operaciones matriciales altamente paralelizadas (Vaswani et al., 2017). A diferencia de los sistemas algorítmicos deterministas tradicionales, donde el comportamiento del sistema puede derivarse lógicamente de sus parámetros de entrada mediante trazabilidad deductiva, los modelos de frontera operan mediante inferencia estocástica sobre distribuciones de probabilidad aprendidas durante el entrenamiento con corpus textuales de escala sin precedentes (Brown et al., 2020). Tal característica arquitectónica implica que los outputs generados no constituyen derivaciones lineales de los inputs, sino productos de procesos probabilísticos cuya génesis permanece distribuida en miles de millones de parámetros interconectados.

A ello se suma que la escala alcanzada por dichos sistemas desafía cualquier analogía con tecnologías precedentes. Los modelos contemporáneos operan con cientos de miles de millones de parámetros entrenados sobre billones de tokens, generando lo que la literatura técnica denomina capacidades emergentes, habilidades no programadas ni previstas por los diseñadores que surgen espontáneamente al superar ciertos umbrales de escala computacional (Wei et al., 2022). Así, tal fenómeno de emergencia constituye una ruptura epistemológica fundamental respecto de paradigmas anteriores de automatización, donde la funcionalidad del sistema podía ser especificada completamente ex ante. En cambio, capacidades como el razonamiento en cadena, la resolución de problemas matemáticos complejos o la generación de código funcional emergen como propiedades no programadas que incluso sorprenden a los equipos de desarrollo (Bommasani et al., 2021).

En esa misma línea, la opacidad epistémica de tales sistemas no constituye una limitación técnica transitoria que pueda superarse mediante avances en interpretabilidad, sino una propiedad estructural derivada de su arquitectura. Los métodos tradicionales de explicabilidad algorítmica, como LIME, SHAP o mapas de atención, presuponen continuidad local y linealidad aproximada en el espacio de decisión, supuestos que colapsan ante arquitecturas cognitivas recursivas y automoduladas, donde las decisiones emergen de dinámicas

de políticas distribuidas y bucles de retroalimentación (Mahrouk, 2025). La investigación más reciente en interpretabilidad demuestra que las técnicas post hoc de explicación ofrecen aproximaciones que no reflejan fielmente los mecanismos internos del modelo, generando lo que se ha denominado paradoja de la caja blanca, explicaciones que proporcionan transparencia superficial sin garantizar comprensión real del proceso decisional (Salvi et al., 2025). La limitación no es meramente cuantitativa en términos de recursos computacionales o técnicas más precisas, sino cualitativa, ya que la propia estructura del procesamiento distribuido impide la reconstrucción de cadenas causales identificables.

Un examen más detallado del funcionamiento de los mecanismos de atención ilustra con precisión la naturaleza de tal opacidad. En una arquitectura de transformador, cada token de entrada interactúa con todos los demás mediante matrices de atención que ponderan dinámicamente la relevancia de cada elemento para la predicción del siguiente. El proceso, replicado a través de decenas o centenas de capas con múltiples cabezas de atención, genera representaciones intermedias de alta dimensionalidad, donde la información relevante para una decisión específica se encuentra distribuida de forma no localizada (Elhage et al., 2021). Cuando el modelo produce un output determinado, no existe, en sentido técnico riguroso, un subconjunto identificable de parámetros o una secuencia de operaciones que pueda señalarse como causa del resultado. La causalidad está distribuida holísticamente en el sistema, resistiendo cualquier intento de descomposición analítica.

A partir de ello se comprende que la arquitectura genera lo que puede describirse como autonomía emergente imprevisible, la capacidad del sistema para producir comportamientos que no pueden anticiparse a partir del conocimiento completo de sus componentes individuales. A diferencia de la autonomía programada de los sistemas expertos tradicionales, donde el comportamiento autónomo opera dentro de parámetros definidos por reglas explícitas, en los modelos de frontera la autonomía surge de la interacción no lineal entre

componentes, generando outputs que pueden sorprender incluso a sus diseñadores (Kenton et al., 2021). Los fenómenos de alucinación, generación de contenido factualmente incorrecto pero sintácticamente coherente, y los comportamientos instrumentales de autopreservación documentados en modelos avanzados constituyen manifestaciones empíricas de tal autonomía emergente que escapa al control directo de los operadores humanos (Amodei et al., 2016).

En paralelo, la ruptura de la linealidad causal opera simultáneamente en dos dimensiones distinguibles analíticamente. En el plano físico, la distribución del procesamiento a través de miles de millones de parámetros interconectados impide identificar una cadena causal localizada entre input y output, la causalidad física existe, pero su estructura distribuida la hace epistémicamente inaccesible (Antamis et al., 2024). En el plano lógico, la naturaleza estocástica del proceso inferencial implica que un mismo input puede generar outputs diferentes en ejecuciones sucesivas, rompiendo la determinabilidad necesaria para reconstruir retrospectivamente el razonamiento del sistema (Lipton, 2018). La doble ruptura, física y lógica, de la linealidad causal constituye el núcleo del problema que enfrentan los sistemas jurídicos, entidades que producen efectos en el mundo sin que pueda trazarse una cadena causal identificable entre una decisión humana específica y el resultado algorítmico.

A ello se añade que la literatura técnica más reciente advierte que los intentos de resolver la opacidad mediante técnicas de explicabilidad post hoc enfrentan limitaciones fundamentales. Los métodos de atribución de características, diseñados originalmente para modelos discriminativos, resultan epistémicamente insuficientes cuando se aplican a sistemas generativos, donde la pregunta explicativa central no es qué características causaron una clasificación, sino qué mecanismos probabilísticos y factores latentes dieron forma a un resultado generativo (Kumarage & Saarela, 2026). La transferencia directa de técnicas desde la IA tradicional hacia los modelos generativos constituye, desde tal perspectiva, un error categorial, se intenta aplicar

herramientas diseñadas para límites de decisión deterministas a sistemas que operan mediante muestreo estocástico sobre espacios latentes de alta dimensionalidad.

Las implicaciones jurídicas de la caracterización técnica precedente resultan profundas. Si la opacidad epistémica constituye una propiedad ontológica intrínseca de los modelos de frontera, no un defecto corregible, entonces cualquier marco regulatorio que presuponga la posibilidad de reconstruir cadenas causales entre diseño humano y resultado algorítmico se apoya en premisas fácticamente falsas. De igual forma, si la autonomía emergente implica que los comportamientos del sistema no pueden anticiparse plenamente a partir de sus especificaciones de diseño, los modelos tradicionales de responsabilidad por producto, que presuponen defectos identificables en la fabricación o el diseño, resultan estructuralmente inadecuados para capturar daños que emergen del funcionamiento correcto pero impredecible del sistema (Teubner, 2018).

Por ese motivo, la distinción entre sistemas algorítmicos deterministas y modelos de frontera con autonomía emergente resulta crucial para delimitar el ámbito de la problemática jurídica aquí abordada. Los sistemas basados en reglas explícitas, *rule based systems*, aunque complejos, mantienen la propiedad de trazabilidad lógica, dado un output, es posible reconstruir la secuencia de reglas aplicadas que lo generaron. Tal trazabilidad permite que el derecho opere con sus categorías tradicionales, imputando el resultado al diseñador que formuló las reglas o al operador que las aplicó incorrectamente (Suksi, 2021). En cambio, los modelos de frontera introducen una ruptura cualitativa, sus outputs emergen de procesos que no pueden ser reconstruidos ni siquiera por sus propios creadores, no por falta de recursos o conocimiento, sino por la estructura misma del procesamiento distribuido.

La caracterización técnica expuesta revela que la autonomía emergente y la opacidad epistémica de los modelos de frontera no constituyen abstracciones teóricas, sino condiciones estructurales que se proyectan directamente sobre la práctica social. Para mostrar cómo

tales propiedades alteran de modo tangible ámbitos donde se deciden derechos fundamentales, resulta necesario examinar su despliegue concreto en sectores jurídicamente sensibles, análisis que se desarrolla en la sección siguiente.

3 IMPACTO DISRUPTIVO EN SECTORES JURÍDICAMENTE SENSIBLES

La caracterización abstracta de las propiedades ontológicas de los modelos de frontera adquiere concreción problemática al examinar su integración efectiva en dominios donde las decisiones algorítmicas producen consecuencias directas sobre derechos fundamentales. Los sectores de salud, administración pública y justicia constituyen ámbitos paradigmáticos donde la tensión entre autonomía emergente algorítmica y garantías jurídicas tradicionales se manifiesta con especial intensidad, revelando la insuficiencia de los marcos regulatorios sectoriales vigentes para contener los riesgos específicos de la IA generativa. El análisis de tales manifestaciones operativas demuestra empíricamente que los estándares tradicionales de *lex artis* profesional y responsabilidad administrativa no logran captar adecuadamente los daños derivados de sistemas cuyo funcionamiento escapa a las categorías de negligencia, impericia o defecto de producto sobre las que se construyó la dogmática de la responsabilidad.

En el ámbito sanitario, el uso de AGI se integra progresivamente en sistemas de apoyo diagnóstico, análisis de historiales clínicos, triaje automatizado y generación de documentación médica. Esa integración introduce una transformación cualitativa respecto de tecnologías médicas precedentes, mientras los sistemas expertos tradicionales operaban como extensiones del juicio clínico humano aplicando reglas diagnósticas codificadas por especialistas, los modelos generativos producen recomendaciones que emergen de procesos inferenciales opacos cuya relación con el conocimiento médico establecido no puede verificarse

directamente (Price, 2018). Cuando un modelo de frontera sugiere un diagnóstico o propone un tratamiento, el profesional sanitario que adopta tal recomendación enfrenta una aporía epistémica, debe evaluar la corrección de una sugerencia cuya génesis no puede reconstruir, sin contar con criterios independientes para distinguir inferencias válidas de alucinaciones plausibles.

A continuación, el problema se intensifica en contextos de urgencia, donde las restricciones temporales impiden realizar una verificación independiente. Los sistemas de triaje automatizado que incorporan modelos generativos para priorizar pacientes según gravedad estimada producen clasificaciones con consecuencias potencialmente letales, sin que sea posible auditar cada decisión en tiempo real. La literatura documenta que los sesgos incorporados en los datos de entrenamiento, reflejo de inequidades históricas en el acceso a servicios sanitarios, se reproducen y amplifican en las recomendaciones algorítmicas, generando discriminación sistemática contra grupos vulnerables bajo apariencia de objetividad técnica (Char et al., 2018). El estándar tradicional de *lex artis* médica, basado en la diligencia profesional evaluable mediante protocolos clínicos establecidos, no ofrece criterios para determinar cuándo la adopción de una recomendación algorítmica opaca constituye conducta diligente o negligencia culpable.

En un plano adicional, el diseño de la arquitectura de responsabilidad en el sector sanitario presupone la identificación de un agente humano cuya conducta pueda evaluarse conforme a estándares de cuidado profesional. Cuando un error diagnóstico resulta de la aplicación incorrecta de conocimiento médico establecido, el sistema jurídico dispone de categorías para imputar responsabilidad al profesional que incurrió en impericia o negligencia. Sin embargo, cuando el error proviene de una recomendación algorítmica que el profesional no estaba en condiciones epistémicas de evaluar críticamente, la cadena de imputación se fragmenta, el médico confió en un sistema cuyo funcionamiento no podía comprender, el desarrollador del modelo no anticipó el comportamiento específico que generó el daño y el

hospital que implementó el sistema cumplió con los estándares técnicos disponibles (Mello & Guha, 2024). La víctima enfrenta lo que la doctrina ha denominado *probatio diabolica*, la imposibilidad práctica de demostrar nexo causal cuando la cadena decisional permanece estructuralmente opaca (Pagallo, 2013).

Por otro lado, en la administración pública, la incorporación de modelos de frontera en servicios gubernamentales introduce tensiones específicas con principios constitucionales que rigen la actuación administrativa. Los sistemas conversacionales que funcionan como asistentes virtuales para trámites, orientación en seguridad social, interpretación normativa o provisión de información no se limitan a transmitir datos predefinidos, sino que generan respuestas contextualizadas que pueden condicionar la conducta ciudadana y anticipar decisiones administrativas sustantivas (Zalnieriute et al., 2019). Cuando un ciudadano recibe orientación algorítmica sobre requisitos para prestaciones sociales, condiciones para permisos o interpretación normativa, esa orientación produce efectos prácticos comparables a actos administrativos informales, sin estar sujeta a controles procedimentales que garanticen la legalidad de la actuación pública.

En consecuencia, el principio de legalidad administrativa exige que toda actuación del poder público se fundamente en normas preexistentes y pueda ser objeto de control jurisdiccional. Los modelos de frontera integrados en servicios públicos generan respuestas que, si bien no constituyen actos administrativos definitivos, condicionan materialmente el ejercicio de derechos. La imposibilidad de reconstruir la *ratio decidendi* de una recomendación algorítmica implica que el ciudadano afectado por una orientación errónea carece de medios epistémicos para impugnar una determinación cuyas razones no pueden identificarse (Citron, 2008). El derecho administrativo tradicional, fundado en la premisa de que las decisiones emanan de funcionarios identificables que aplican normas conocidas mediante razonamientos reconstruibles, no dispone de categorías para abordar actuaciones cuya génesis permanece distribuida en sistemas algorítmicos opacos.

Asimismo, la experiencia comparada ofrece casos paradigmáticos donde la automatización administrativa sin salvaguardas adecuadas provocó vulneraciones masivas de derechos. El sistema australiano de recuperación automatizada de deudas sociales, Robodebt, muestra cómo la eliminación de la revisión humana y la discrecionalidad administrativa puede generar injusticias sistemáticas cuando los algoritmos operan sobre premisas erróneas sin posibilidad de corrección individual (Priergaard, 2024). Aunque se trataba de sistemas deterministas más que generativos, anticipa problemáticas que se amplifican con modelos cuya opacidad deriva de la estructura misma del procesamiento algorítmico. La transición desde sistemas basados en reglas hacia modelos de frontera intensifica exponencialmente los riesgos de actuaciones lesivas que no pueden ser contenidas por mecanismos tradicionales del derecho administrativo.

En el ámbito judicial, la incorporación de modelos de lenguaje en la elaboración de escritos, análisis preliminar de casos, investigación jurisprudencial o redacción de resoluciones afecta directamente la función jurisdiccional del Estado. La introducción de entidades cuyo funcionamiento escapa al control racional que legitima la autoridad judicial plantea desafíos sustanciales (Remus & Levy, 2017). El caso *Mata v. Avianca*, en que un abogado presentó precedentes fabricados por ChatGPT como fundamento de su escrito, ofrece evidencia empírica de los riesgos de erosión verificatoria que genera la delegación de funciones cognitivas críticas a sistemas cuyas limitaciones epistémicas no se comprenden adecuadamente (Tritto & Torres, 2025).

En una línea convergente, el problema no se reduce a casos de uso indebido por operadores individuales, sino que afecta la estructura misma del razonamiento judicial. Cuando un modelo genera argumentos que luego son adoptados por jueces o abogados, se produce una transferencia de autoría intelectual con implicaciones no resueltas para la legitimidad del acto jurídico. El argumento jurídico presupone un autor responsable epistémicamente de sus afirmaciones, susceptible de ser interpelado por sus fundamentos. Los modelos generativos, en

cambio, producen texto con forma argumentativa pero sin sustancia reflexiva (Srivastava, 2025). La diferencia ontológica entre el procesamiento probabilístico de patrones lingüísticos y el razonamiento jurídico genuino, que implica comprensión normativa, ponderación de valores y responsabilidad sobre las conclusiones, no puede salvarse mediante mejoras técnicas en la capacidad generativa del modelo.

Los sistemas predictivos que utilizan algoritmos para evaluar riesgo de reincidencia, determinar condiciones de libertad provisional o informar decisiones de sentencia introducen sesgos sistemáticos derivados de datos históricos que reflejan discriminaciones estructurales en el sistema penal (O'Neil, 2016). El caso COMPAS, en Estados Unidos, evidenció cómo algoritmos con apariencia de objetividad estadística reproducían y amplificaban sesgos raciales, clasificando a personas afroamericanas como de mayor riesgo independientemente de sus circunstancias individuales (Barocas & Selbst, 2016). Aunque se trataba de sistemas previos al auge de los modelos generativos, las problemáticas allí reveladas se intensifican cuando la opacidad impide identificar incluso qué variables resultan determinantes para las predicciones producidas.

Las manifestaciones sectoriales examinadas evidencian que la integración de sistemas opacos y autónomos no solo genera daños puntuales, sino que tensiona los fundamentos mismos sobre los que se articula la juridicidad en contextos constitucionales. Precisamente por ello, el paso siguiente consiste en analizar cómo tales transformaciones erosionan de manera estructural las garantías que sostienen el Estado de Derecho, cuestión que constituye el foco de la sección posterior.

4 EROSIÓN ESTRUCTURAL DE LAS GARANTÍAS CONSTITUCIONALES

La integración de sistemas algorítmicos opacos en procesos decisionales que afectan derechos fundamentales produce una lesión

estructural a los pilares del Estado Constitucional, una afectación que trasciende las problemáticas sectoriales analizadas en el apartado precedente. Los principios de motivación de las decisiones públicas, contradicción efectiva y tutela judicial constituyen garantías interconectadas cuya función no se reduce a un requisito procedimental, sino que delimita la legitimidad del acto jurídico en el Estado de Derecho. La opacidad epistémica de los modelos de frontera no representa un obstáculo práctico que pudiera superarse mediante esfuerzos de transparencia, sino una condición que anula la posibilidad lógica del control jurisdiccional y desarticula los fundamentos sobre los que descansa la legitimidad democrática del ejercicio del poder público.

Desde esta perspectiva, el principio de motivación de las decisiones que afectan derechos subjetivos constituye una conquista central del constitucionalismo moderno. Su función supera la formalidad procedimental y opera como garantía sustantiva contra la arbitrariedad del poder (Alexy, 1989). La exigencia de que toda determinación pública que incida sobre la esfera jurídica de los ciudadanos exprese las razones que la fundamentan cumple una triple función: permite al destinatario comprender por qué se decidió en determinado sentido, habilita la impugnación mediante argumentos que controviertan esas razones y posibilita el control jurisdiccional sobre la corrección del razonamiento expresado (Atienza, 2013). Las tres funciones presuponen la existencia de una *ratio decidendi* identificable, un itinerario argumentativo que conecte premisas fácticas y normativas con la conclusión decisoria, susceptible de reconstrucción, evaluación y controversia.

A partir de allí, la incorporación de agentes inteligentes en procesos decisionales jurídicamente relevantes erosiona dicho principio de manera estructural al introducir una capa de mediación algorítmica cuyas operaciones internas no constituyen razones en el sentido que el derecho constitucional presupone. Cuando un sistema genera una recomendación, clasificación o determinación, no existe, en sentido técnicamente riguroso, una cadena de razones reconstruible que permita comprender por qué se llegó a un resultado específico y no a otro

(Wachter et al., 2017). Los *outputs* algorítmicos emergen de procesos de inferencia distribuida sobre espacios de alta dimensionalidad, sin que pueda identificarse un conjunto de premisas cuya concatenación lógica produzca la conclusión. La opacidad no es un defecto corregible sino una propiedad definitoria, pues el sistema no razona aplicando reglas a hechos para derivar conclusiones, sino que genera predicciones probabilísticas cuya génesis permanece distribuida en miles de millones de parámetros interconectados.

Desde luego, tal característica produce una incompatibilidad estructural con la función que el principio de motivación cumple en el Estado de Derecho. Si la legitimidad de una decisión que afecta derechos depende de la existencia de razones controlables, las determinaciones cuya génesis permanece opaca carecen de las condiciones mínimas de legitimidad que el constitucionalismo exige (Ferrajoli, 2011). El problema no radica en razones insuficientes, erróneas o mal expresadas, vicios corregibles mediante recursos ordinarios, sino en que la estructura misma del proceso decisional impide la existencia de razones en el sentido jurídicamente relevante. La decisión algorítmica adopta la forma externa de una determinación, pero carece de la sustancia argumentativa que la convertiría en acto jurídico legítimo.

Por otro lado, el principio de contradicción, garantía procesal derivada del derecho de defensa, presupone también condiciones epistémicas que la opacidad algorítmica anula. Para que la contradicción sea efectiva, el destinatario de una decisión adversa debe conocer sus fundamentos y formular argumentos que los controviertan; el órgano decisor debe evaluar esos argumentos y, en su caso, modificar la determinación inicial (Guastini, 2014). Cuando los fundamentos de una decisión permanecen opacos, no por ocultamiento deliberado sino por imposibilidad estructural de reconstrucción, la contradicción se convierte en ejercicio formal vacío de contenido: el afectado puede expresar desacuerdo genérico, pero no identificar premisas específicas que impugnar ni argumentos relevantes para modificar el resultado.

De manera correlativa, el ciudadano afectado por una determinación algorítmica adversa enfrenta una aporía constitucional, dado que el ordenamiento le reconoce derecho a contradecir mientras la estructura del proceso decisional impide el ejercicio efectivo de ese derecho. La asimetría epistémica entre el sistema algorítmico, que procesa información mediante patrones inaccesibles, y el destinatario, que solo conoce el resultado sin acceso a su génesis, produce una vulneración del principio de igualdad de armas que no puede corregirse mediante garantías procedimentales adicionales (Hildebrandt, 2016). La desigualdad no deriva de restricciones de acceso a información subsanables, sino de la naturaleza misma del proceso: incluso con acceso completo a los parámetros del modelo, el destinatario carecería de medios para traducir esa información técnica en argumentos jurídicamente relevantes para su defensa.

En una dimensión complementaria, la garantía de tutela judicial efectiva completa el sistema de garantías constitucionales afectado por la opacidad algorítmica. Para que el control jurisdiccional opere como límite efectivo al poder, los tribunales deben evaluar la corrección de las decisiones sometidas a su conocimiento mediante criterios jurídicos articulables (Ferrajoli, 2011). Dicha evaluación presupone que las decisiones controladas expresan razones contrastables con el ordenamiento aplicable, de modo que el tribunal verifique si los hechos fueron correctamente establecidos, si las normas fueron interpretadas de manera adecuada y si la conexión entre premisas y conclusión resulta lógicamente válida. Cuando las decisiones emergen de procesos opacos que no expresan razones reconstruibles, el control jurisdiccional pierde su objeto, ya que no hay razonamiento que evaluar, solo resultados que constatar.

A modo de ilustración, la transformación cualitativa que se describe puede observarse con claridad al contrastarla con formas tradicionales de discrecionalidad administrativa. Cuando un funcionario ejerce potestades discrecionales, el control jurisdiccional opera sobre los límites de dicha discrecionalidad: verifica que la decisión

respete principios de razonabilidad, proporcionalidad y finalidad pública, sin sustituir el juicio administrativo por el judicial (Paredes, 2020). Ese control presupone un juicio administrativo expresable mediante razones, incluso si el ordenamiento reconoce márgenes de apreciación al órgano decisor. La opacidad algorítmica produce un escenario distinto, dado que no hay juicio cuya razonabilidad evaluar, sino un proceso computacional cuya corrección solo podría verificarse mediante criterios técnicos ajenos a las categorías del control jurídico.

Frente a este panorama, los tribunales enfrentan un dilema estructural cuando conocen impugnaciones contra decisiones con componentes algorítmicos opacos. Pueden optar por un control formal que verifique únicamente requisitos procedimentales externos, como la autorización del sistema, la notificación del resultado o la indicación de vías de recurso, renunciando a cualquier evaluación sustantiva del proceso. En alternativa, pueden intentar un control material que exija traducir el proceso algorítmico a términos jurídicamente evaluables, aunque dicha traducción resulte técnicamente imposible cuando la opacidad es irreductible. Ninguna opción resulta satisfactoria para el Estado de Derecho: la primera vacía de contenido la tutela judicial y la segunda impone exigencias que la arquitectura técnica de los sistemas no puede satisfacer.

Por tanto, la crisis constitucional diagnosticada no afecta solo a casos donde los modelos de frontera adoptan decisiones finales sin intervención humana, sino que se extiende a situaciones en las que condicionan materialmente el contenido de decisiones formalmente adoptadas por personas. Cuando un funcionario, juez o profesional adopta una determinación basándose de modo sustancial en recomendaciones algorítmicas cuya génesis no puede evaluar críticamente, la motivación formal expresada no refleja el proceso real de formación de la decisión. Las razones expuestas constituyen una racionalización posterior de un resultado previamente generado por el sistema, no una justificación genuina del acto decisorio (Elish, 2019). La forma

externa de la motivación se preserva, mientras su función constitucional, permitir contradicción y control, se vacía de contenido.

Según lo ha señalado una parte de la literatura especializada, la situación descrita ha sido caracterizada como tránsito desde el Estado de Derecho hacia el Estado de Algoritmo, un régimen donde las determinaciones que afectan derechos no emergen de la aplicación controlable de normas jurídicas preexistentes, sino de procesos computacionales cuya lógica escapa a las categorías del derecho (Suksi, 2021). Tal caracterización no implica que los sistemas algorítmicos operen fuera del derecho formal, puesto que siguen existiendo normas que autorizan su uso y regulan sus efectos, sino que la función del derecho como límite del poder mediante razones públicamente controlables queda estructuralmente comprometida. El derecho puede seguir nominando categorías y prescribiendo consecuencias, aunque pierde capacidad para operar como freno efectivo frente a decisiones cuya génesis permanece opaca.

La erosión constitucional identificada muestra que la opacidad algorítmica imposibilita la reconstrucción de razones y la verificación de procesos, comprometiendo las condiciones epistémicas indispensables para la legitimidad pública. Ante esa crisis, los marcos regulatorios han intentado preservar la imputación mediante el paradigma del control humano significativo, cuya aparente solución demanda un examen crítico específico que se aborda en la sección siguiente.

5 INVIABILIDAD ESTRUCTURAL DEL PARADIGMA DE SUPERVISIÓN HUMANA

Frente a las tensiones documentadas en los apartados precedentes, los marcos regulatorios contemporáneos han promovido el paradigma del Control Humano Significativo (MHC) como principio rector para la gobernanza de sistemas de IA de alto riesgo. La noción, originada en el debate sobre sistemas de armas autónomas letales y

luego extendida a aplicaciones civiles, postula que la responsabilidad humana sobre decisiones algorítmicas puede preservarse mediante mecanismos de supervisión que mantengan a personas físicas en posición de controlar efectivamente los resultados del sistema (Santoni de Sio & van den Hoven, 2018).

El Reglamento Europeo de IA (2024/1689) establece en su artículo 14 que los sistemas de alto riesgo deben incorporar mecanismos que aseguren una supervisión humana efectiva, lo que implica diseñarlos con interfaces adecuadas, garantizar la posibilidad de interpretar sus salidas, permitir su interrupción segura y adoptar medidas proporcionales al nivel de autonomía y al riesgo. Sin embargo, un examen riguroso de las condiciones cognitivas y temporales que el MHC presupone revela su inviabilidad estructural frente a la velocidad, escala y opacidad propias de los AGI, lo que expone que su exigencia legislativa constituye una ficción jurídica que no logra restituir la imputación real de responsabilidad.

Desde un enfoque conceptual, el marco filosófico del MHC se articula convencionalmente mediante dos condiciones necesarias: rastreo (*tracking*) y trazabilidad (*tracing*). La condición de rastreo exige que el sistema responda a las razones morales relevantes de los agentes humanos implicados, mientras que la de trazabilidad requiere que cualquier resultado del sistema pueda vincularse a al menos un agente humano con capacidad técnica para comprender su funcionamiento y disposición moral para asumir responsabilidad por sus consecuencias (Santoni de Sio et al., 2023). Formuladas inicialmente en el contexto de sistemas de armas autónomas, donde las decisiones conllevan consecuencias irreversibles, ambas condiciones se han propuesto como estándar general para cualquier sistema algorítmico que afecte derechos o intereses legítimos.

Al trasladar dicho marco a los AGI, emergen obstáculos de naturaleza estructural, no meramente práctica. La condición de rastreo presupone que es posible verificar si el sistema opera conforme a las razones del agente humano, lo cual requiere acceso epistémico al

proceso decisional para contrastarlo con las motivaciones que deberían guiarlo. En modelos deterministas basados en reglas explícitas, tal verificación resulta viable en principio, ya que puede examinarse si las reglas codificadas reflejan adecuadamente las razones normativas pertinentes y si el sistema las aplica correctamente. No ocurre lo mismo en los modelos de frontera, cuya opacidad epistémica impide tal verificación, pues ni siquiera con conocimiento completo de los parámetros resulta posible determinar si un resultado específico responde a razones morales del operador o si surge de patrones estadísticos ajenos a dichas motivaciones (Veluwenkamp, 2022).

Con base en aportes recientes de la filosofía, se ha identificado una dificultad adicional que afecta incluso a configuraciones ideales donde pudiera superarse la opacidad: la subdeterminación de razones (*reasons underdetermination*). Las razones morales relevantes para una decisión no siempre conducen a un único curso de acción correcto, sino que pueden permitir múltiples opciones igualmente compatibles con tales razones (Kozlovski, 2025). Cuando el sistema selecciona una opción entre varias aceptables, no puede afirmarse con certeza que dicha selección responda a razones del agente humano ni que las contradiga, ya que la relación entre razones y resultado permanece indeterminada. Esta indeterminación impide cumplir plenamente la condición de rastreo incluso si no existiera opacidad técnica, pues las propias razones no especifican con precisión qué conducta debe adoptar el sistema en contextos con alternativas igualmente justificables.

En paralelo, la condición de trazabilidad enfrenta dificultades semejantes. Vincular resultados algorítmicos a agentes humanos responsables implica una estructura causal en la que decisiones de diseño, configuración u operación determinan comportamientos del sistema de manera que permitan atribuir consecuencias específicas. Sin embargo, los modelos de frontera, caracterizados por la emergencia de comportamientos no anticipados a partir de procesos de entrenamiento no supervisados directamente, disuelven dicha estructura causal, ya

que los resultados no derivan de decisiones humanas individuales sino que emergen de interacciones no lineales entre componentes cuyo comportamiento agregado no fue diseñado ni anticipado (Teubner, 2018). La responsabilidad, en consecuencia, se diluye entre una cadena de actores que incluye a investigadores, entrenadores, implementadores y operadores, sin que ninguno pueda ser identificado como causa determinante de un resultado concreto.

Por otro lado, las condiciones cognitivas necesarias para la supervisión significativa se presentan como especialmente problemáticas. Para que exista un control humano real, el operador debe ser capaz de evaluar críticamente las recomendaciones, detectar errores potenciales e intervenir de forma adecuada. Tal capacidad exige una comprensión suficiente del funcionamiento interno del sistema como para formular juicios independientes respecto de la corrección de sus resultados. En los modelos de frontera, esa comprensión es inalcanzable, incluso para sus propios diseñadores, lo que impide que operadores humanos evalúen de manera informada los outputs generados. Así, la supervisión queda reducida a una validación meramente formal, donde los operadores aprueban o rechazan resultados sin disponer de criterios que permitan distinguir con certeza entre errores y respuestas plausibles pero incorrectas (Green & Chen, 2019).

Además, la psicología cognitiva ha documentado sesgos estables que erosionan la eficacia de la supervisión, incluso en condiciones donde parecería factible. El sesgo de automatización, consistente en confiar en exceso en recomendaciones generadas automáticamente, se intensifica cuando los operadores carecen de tiempo o recursos para realizar una verificación independiente, justo el tipo de contexto donde suelen operar los modelos de frontera (Parasuraman & Riley, 1997). Junto a ello, la complacencia algorítmica, fenómeno que debilita la vigilancia crítica después de periodos de aparente buen desempeño, reduce progresivamente la atención incluso de operadores originalmente diligentes. Estos sesgos no son anomalías corregibles mediante capacitación, sino patrones persistentes del funcionamiento

mental humano que deben ser tomados en cuenta en cualquier diseño institucional.

También las condiciones temporales resultan incompatibles con los requisitos del MHC. Los modelos de frontera operan a velocidades que superan por mucho la capacidad de revisión humana, generando millones de outputs en fracciones de segundo (Crootof, 2016). Garantizar supervisión efectiva exigiría ralentizar el funcionamiento del sistema hasta hacerlo compatible con los ritmos de procesamiento humano, lo cual eliminaría precisamente las ventajas que justifican su uso. Como alternativa, se ha propuesto la supervisión por muestreo, es decir, revisar aleatoriamente una fracción de las decisiones, pero tal enfoque no satisface la exigencia de control significativo: los resultados no revisados permanecen sin supervisión real y, una vez ejecutadas, las decisiones ya no pueden ser corregidas aunque se identifiquen errores graves.

Observaciones empíricas en sectores concretos confirman la imposibilidad práctica de satisfacer el MHC como garantía de responsabilidad. En el ámbito marítimo, por ejemplo, los sistemas de navegación autónoma que emplean IA para procesar el entorno y tomar decisiones operan en contextos donde la supervisión humana resulta inviable. La velocidad de los eventos, la complejidad del entorno físico y la opacidad de los procesos internos imposibilitan que los operadores ejerzan un control efectivo sobre decisiones críticas (Albrecht et al., 2025). En tales escenarios, el MHC no funciona como una condición técnica alcanzable, sino como ideal regulativo que puede orientar el diseño futuro sin resolver los déficits actuales de imputación.

En términos normativos, la incorporación del MHC en marcos jurídicos como el Reglamento Europeo de IA da lugar a una paradoja. El derecho impone la exigencia de supervisión significativa, aunque los sistemas regulados no puedan satisfacerla materialmente. Así, la exigencia se transforma en ficción jurídica: se presume la existencia de control humano simplemente porque la norma lo establece, sin que ello se corresponda con condiciones técnicas reales. Tal ficción

cumple una función de legitimación simbólica, ya que permite sostener la narrativa de control humano sobre sistemas que, en la práctica, operan más allá de dicho control (Yeung, 2017).

Las implicaciones de esta ficción son especialmente graves en el plano de la responsabilidad jurídica. En caso de que el MHC funcionara como se espera, sería posible conservar los esquemas tradicionales de imputación, permitiendo atribuir daños a quienes ejercieran mal la supervisión, conforme a estándares de negligencia. Pero si el control humano nunca fue real, no cabe reprochar su ausencia a quienes carecían de condiciones para ejercerlo. Y, al mismo tiempo, el marco normativo impide reconocer autonomía al sistema, pues la ley sigue presuponiendo control humano donde no lo hay. El resultado es un vacío de responsabilidad: no puede imputarse a quien no controlaba, ni reconocerse como sujeto a un sistema que, formalmente, no debía actuar por sí solo.

Desde la teoría crítica, la noción de zona de deformación moral (*moral crumple zone*) permite describir con precisión esta situación: entornos donde los operadores humanos aparecen como responsables de fallos que no tienen capacidad real de evitar, cargando con consecuencias que deberían distribuirse entre múltiples actores o reconducirse al diseño sistémico (Elish, 2019). Bajo su formulación vigente, el MHC no soluciona esta dinámica, sino que la acentúa. Exige supervisión sin garantizar medios para ejercerla, crea figuras de responsabilidad formal carentes de poder real, y convierte a los operadores en chivos expiatorios de fallos cuya causa efectiva se encuentra aguas arriba, en decisiones estructurales de diseño y despliegue que las organizaciones responsables eluden sistemáticamente.

La inviabilidad estructural del CHS deja sin anclaje real los esquemas tradicionales de imputación, generando un vacío de responsabilidad imposible de resolver con categorías clásicas. Superado el paradigma supervisorial, la discusión debe desplazarse hacia alternativas dogmáticas capaces de rearticular la atribución de consecuencias en entornos de agencia no humana, cuestión que conduce directamente a

la fundamentación de la personalidad jurídica algorítmica desarrollada en la sección siguiente.

6 FUNDAMENTACIÓN DOGMÁTICA DE LA PERSONALIDAD JURÍDICA ALGORÍTMICA

El diagnóstico desarrollado en los apartados precedentes conduce a una conclusión ineludible, las categorías jurídicas tradicionales resultan estructuralmente inadecuadas para someter a categorías de imputación la agencia funcional de los AGI. La opacidad epistémica impide satisfacer las exigencias constitucionales de motivación y control, la autonomía emergente desborda los marcos sectoriales de responsabilidad profesional, el paradigma del control humano significativo opera como ficción inoperante que no restituye imputación real. Se configura así un vacío normativo donde los daños causados por sistemas algorítmicos permanecen sin responsable identificable, vulnerando tanto los derechos de las víctimas a obtener reparación como la función ordenadora del derecho mediante la atribución de consecuencias a conductas. En ese escenario, la creación de una personalidad jurídica algorítmica como *tertium genus* funcional emerge no como opción teórica entre otras disponibles, sino como mecanismo dogmáticamente coherente para restablecer nexos de imputación y patrimonio en un entorno donde la agencia humana directa ha desaparecido fácticamente.

Para desarrollar dicha propuesta, resulta imprescindible situar el debate en sus coordenadas teóricas adecuadas. La personalidad jurídica no constituye un atributo natural de las entidades, sino una construcción normativa mediante la cual el ordenamiento selecciona centros de imputación de derechos y obligaciones. A partir de las formulaciones clásicas de la dogmática civilista, la capacidad jurídica se concibe como técnica destinada a identificar puntos de atribución de consecuencias jurídicas, sin importar si corresponden a realidades

biológicas o a construcciones sociales (Kelsen, 1982). La extensión histórica de la personalidad a corporaciones, fundaciones, asociaciones y entes estatales demuestra, en efecto, la flexibilidad conceptual del derecho para reconocer como sujetos a entidades carentes de voluntad en sentido psicológico, siempre que existan razones funcionales que justifiquen tal atribución (Gierke, 2001).

En consonancia con lo anterior, la teoría de sistemas sociales proporciona un marco conceptual especialmente idóneo para comprender tanto la naturaleza del problema como los fundamentos de la solución propuesta. Desde dicha perspectiva, los sistemas sociales operan mediante comunicaciones, no a través de acciones individuales. La atribución de acción a personas representa una simplificación funcional adoptada por el sistema jurídico, pero no refleja necesariamente la estructura real de los procesos comunicativos (Luhmann, 1995). Los AGI participan activamente en la comunicación social, generan textos con efectos performativos, responden a interpelaciones y modifican expectativas de otros actores. Lo relevante, desde una visión sistémica, no es si tales sistemas poseen pensamiento o conciencia, categorías antropocéntricas inaplicables, sino si generan diferencias operativas con efectos en sistemas funcionales (Teubner, 2018).

En consecuencia, la pregunta jurídicamente pertinente no se refiere a la capacidad de pensar de las máquinas, sino a su capacidad de actuar en el sentido normativo de producir efectos que demandan atribución (Teubner, 2018). La evidencia empírica indica que, efectivamente, los modelos de frontera generan recomendaciones que condicionan decisiones humanas, producen textos con consecuencias jurídicas, clasifican personas para otorgar servicios, negocian condiciones contractuales y ejecutan transacciones. Tales actividades producen efectos jurídicos, como la generación de obligaciones o la causación de daños, independientemente de la falta de intencionalidad fenomenológica. Así como el derecho ha reconocido personalidad a corporaciones sin mente por su capacidad de producir efectos, se enfrenta ahora a

entidades algorítmicas que también los generan, aunque aún carecen de reconocimiento como centros de imputación.

Dentro del marco conceptual descrito, la propuesta de personalidad jurídica algorítmica requiere diferenciarse cuidadosamente de visiones que antropomorfizan la tecnología o que buscan atribuir derechos subjetivos a las máquinas. El objetivo no es reconocer dignidad, conciencia ni intereses protegibles a los modelos de frontera, sino establecer una categoría funcional que permita operar jurídicamente sobre entidades que ya producen efectos relevantes, pero que escapan a las estructuras de imputación disponibles. La personalidad sugerida es de carácter instrumental y alcance limitado, no comporta el reconocimiento de derechos fundamentales ni de capacidad general, sino la habilitación específica para que tales entidades sean consideradas centros de imputación de obligaciones y responsabilidades en los contextos donde operan de forma autónoma (Solum, 1992).

Conviene destacar que el modelo del *peculium* romano, rescatado por la literatura jurídica contemporánea, ofrece un antecedente dogmático de notable valor. En el derecho romano, los esclavos no eran sujetos de derecho, pero podían administrar un patrimonio separado (*peculium*) que respondía por las obligaciones contraídas en el tráfico económico. Ese mecanismo limitaba la responsabilidad del amo a dicho patrimonio y garantizaba solvencia en la actividad comercial del esclavo (Teubner, 2018). Tal solución permitió abordar una forma de agencia que no encajaba en la dicotomía clásica persona/cosa, mediante la creación de un *tertium genus* funcional que facilitaba las operaciones jurídicas sin modificar la teoría general. La personalidad jurídica algorítmica podría operar de manera análoga, con un patrimonio afectado constituido por seguros obligatorios, fondos de garantía o mecanismos de capitalización directa, destinado a responder por los daños derivados de la operación autónoma. Ello no implicaría atribuir personalidad plena al sistema ni exonerar completamente a los humanos involucrados en su diseño y despliegue.

Para que esta arquitectura resulte operativa, sería necesario articular un conjunto de mecanismos normativos interrelacionados. En primer lugar, se requeriría el reconocimiento jurídico de modelos de frontera que superen ciertos umbrales de autonomía y opacidad como sujetos limitados, inscritos en un registro y dotados de identificación única que permita su trazabilidad. En segundo lugar, se establecería la obligación de constituir un patrimonio de respaldo, mediante seguros, depósitos o fondos específicos, proporcional al riesgo asociado a su funcionamiento y destinado exclusivamente a la satisfacción de responsabilidades. En tercer lugar, se adoptarían reglas que atribuyan al sistema, como centro primario de imputación, las consecuencias derivadas de su operación autónoma, sin perjuicio de la eventual concurrencia de responsabilidad de diseñadores, operadores o beneficiarios por sus propias acciones (Čerka et al., 2015).

Desde una perspectiva comparativa, esta arquitectura presenta ventajas dogmáticas relevantes frente a las alternativas actualmente en discusión. A diferencia de la extensión de la responsabilidad objetiva a desarrolladores u operadores, que puede generar cargas excesivas sobre actores sin control efectivo del comportamiento del sistema e inhibir la innovación útil, la personalidad algorítmica concentra la imputación en el origen del riesgo, sin presumir un control que no existe (Scherer, 2016). En contraste con los esquemas de socialización del riesgo mediante fondos públicos de compensación, que reducen los incentivos preventivos al desvincular los costos del daño de las actividades que los generan, la personalidad algorítmica mantiene el vínculo entre sistema causante y patrimonio resarcitorio. Frente a la ficción del control humano significativo, que exige una supervisión técnicamente inviable, esta propuesta reconoce la autonomía operativa de los sistemas y construye mecanismos jurídicos acordes con dicha realidad.

Una de las objeciones más recurrentes frente a esta figura apunta al riesgo de que funcione como instrumento de exoneración para los humanos implicados en la creación y despliegue de sistemas dañinos. Si

bien se trata de una crítica legítima, parte de una presuposición equivocada, al asumir que el reconocimiento de personalidad algorítmica implica necesariamente exclusión de responsabilidad humana. En realidad, el diseño propuesto opera en sentido contrario. Al establecer un centro de imputación específico para la operación autónoma, permite delimitar con mayor precisión las responsabilidades diferenciadas de diseñadores por fallos estructurales, entrenadores por sesgos de datos, desplegadores por configuraciones inapropiadas y operadores por usos indebidos, sin forzar la atribución íntegra de la responsabilidad a sujetos que carecían de control sobre la conducta que finalmente produjo el daño (Pagallo, 2013).

7 CONCLUSIONES

La interacción entre los AGI y el ordenamiento jurídico no debe interpretarse como una mera fricción regulatoria coyuntural, sino como una disrupción ontológica que desafía los presupuestos de la imputación normativa. El análisis realizado sugiere que la arquitectura estocástica y la autonomía emergente de estos sistemas generan una ruptura estructural en la linealidad causal física y lógica sobre la que se asienta el Estado constitucional. Al operar mediante procesos de inferencia distribuida que resultan epistémicamente opacos incluso para sus creadores, estos modelos desarticulan la capacidad del derecho para trazar vínculos fácticos inequívocos entre la intención humana y el resultado algorítmico, lo que erosiona las garantías de motivación y control jurisdiccional necesarias para la legitimidad de las decisiones públicas.

Desde una perspectiva dogmática, se sostiene que el paradigma regulatorio vigente, centrado en el "Control Humano Significativo", ha derivado en una ficción jurídica inoperante frente a la realidad técnica de la IA generativa. La investigación revela que las limitaciones cognitivas y temporales de los operadores humanos hacen inviable una

supervisión efectiva sobre sistemas de "caja negra", transformando la exigencia de control en un mecanismo que, lejos de garantizar la seguridad, crea zonas de "deformación moral" donde la responsabilidad se asigna injustamente a agentes humanos carentes de agencia real. Esta crítica teórica permite reconceptualizar la opacidad no como un defecto técnico subsanable, sino como una propiedad constitutiva que invalida los esquemas tradicionales de responsabilidad subjetiva y objetiva.

Para superar el vacío de imputación identificado, se propone la configuración de una personalidad jurídica algorítmica entendida estrictamente como un *tertium genus* funcional e instrumental, despojado de connotaciones antropocéntricas o morales. Esta categoría, inspirada en figuras como el *peculium* romano, permitiría vincular la agencia autónoma del sistema a un patrimonio de afectación específico, que podría ser constituido mediante seguros o fondos de garantía, para responder por los daños emergentes de su operación. Tal mecanismo ofrece una vía pragmática para restablecer la solvencia del sistema de responsabilidad civil y administrativa, permitiendo que el derecho procese las consecuencias de la autonomía algorítmica sin necesidad de recurrir a ficciones de control humano ni de diluir la reparación de las víctimas en la inescrutabilidad técnica.

El reconocimiento de esta nueva categoría jurídica abre una agenda de investigación expansiva orientada a definir los contornos procesales y económicos de la entidad algorítmica. Resultará imperativo examinar los mecanismos de constitución y gestión de los fondos de riesgo asociados, así como los criterios para determinar el umbral de autonomía que justificaría la atribución de personalidad diferenciada. Asimismo, se requiere un análisis profundo sobre la interacción entre esta responsabilidad patrimonial autónoma y la responsabilidad concurrente de desarrolladores y proveedores, especialmente en lo referente a la distribución de cargas probatorias y la definición de estándares de diligencia en el diseño de arquitecturas de seguridad.

En ese sentido, la resistencia a reconocer subjetividad jurídica a entes no humanos parece responder más a inercias culturales que a

imposibilidades dogmáticas. El derecho, como tecnología de organización social, ha demostrado históricamente su capacidad para atribuir personalidad a entidades abstractas cuando la necesidad funcional lo requiere. Al admitir la posibilidad de una personalidad algorítmica, el sistema jurídico no abdica de su función humanista, sino que la reafirma: adapta sus herramientas formales para asegurar que, en un entorno de agencias distribuidas y poshumanas, prevalezca la protección efectiva de los derechos y la integridad del Estado de Derecho.

REFERENCIAS

ALBRECHT, L.; BRAUN, H.; KOSACK, T. R.; KRÜGER, T. The Limits of Meaningful Human Control of AI in the Maritime Domain. **Transactions on Maritime Science**, v. 14, n. 3, 2025.

ALEXY, R. **Teoría de la argumentación jurídica: La teoría del discurso racional como teoría de la fundamentación jurídica.** Madrid: Centro de Estudios Constitucionales, 1989.

AMODEI, D.; OLAH, C.; STEINHARDT, J.; CHRISTIANO, P.; SCHULMAN, J.; MANÉ, D. Concrete Problems in AI Safety. **arXiv**, 2016.

ANTAMIS, T.; DROSOU, A.; VAFEIADIS, T.; NIZAMIS, A.; IOANNIDIS, D.; TZOVARAS, D. Interpretability of Deep Neural Networks: A Review of Methods, Classification and Hardware. **Neurocomputing**, v. 601, artículo 128204, oct. 2024.

ATIENZA, M. **Curso de argumentación jurídica.** Madrid: Editorial Trotta, 2013.

BAROCAS, S.; SELBST, A. D. Big Data's Disparate Impact. **California Law Review**, v. 104, n. 3, p. 671–732, 2016.

BOMMASANI, R.; HUDSON, D. A.; ADELI, E.; ALTMAN, R.; ARORA, S.; VON ARX, S.; BERNSTEIN, M. S. et al. On the Opportunities and Risks of Foundation Models. **arXiv**, 2021.

BROWN, T. B.; MANN, B.; RYDER, N.; SUBBIAH, M.; KAPLAN, J.; DHARIWAL, P.; NEELAKANTAN, A. et al. Language Models Are

Few Shot Learners. In: **Advances in Neural Information Processing Systems 33**, 2020.

ČERKA, P.; GRIGIENĖ, J.; SIRBIKYTĖ, G. Liability for Damages Caused by Artificial Intelligence. **Computer Law & Security Review**, v. 31, n. 3, p. 376–389, 2015.

CHAR, D. S.; SHAH, N. H.; MAGNUS, D. Implementing Machine Learning in Health Care, Addressing Ethical Challenges. **New England Journal of Medicine**, v. 378, n. 11, p. 981–983, 2018.

CITRON, D. K. Technological Due Process. **Washington University Law Review**, v. 85, n. 6, p. 1249–1313, 2008.

CROOTOF, R. War Torts: Accountability for Autonomous Weapons. **University of Pennsylvania Law Review**, v. 164, n. 6, p. 1347–1402, 2016.

ELHAGE, N.; NANDA, N.; OLSSON, C.; HENIGHAN, T.; JOSEPH, N.; MANN, B.; ASKELL, A. et al. A Mathematical Framework for Transformer Circuits. **Transformer Circuits Thread**, 2021.

ELISH, M. C. Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction. **Engaging Science, Technology, and Society**, v. 5, p. 40–60, 2019.

FERRAJOLI, L. **Principia iuris: Teoría del derecho y de la democracia.** Vol. 1, Teoría del derecho. Madrid: Editorial Trotta, 2011.

GIERKE, O. F. von; TROELTSCH, E. **Natural Law and the Theory of Society, 1500 to 1800.** Traducido por Sir Ernest Barker. Reimpresa. The Lawbook Exchange, Ltd., 2001.

GREEN, B.; CHEN, Y. The Principles and Limits of Algorithm-in-the--Loop Decision Making. **Proceedings of the ACM on Human-Computer Interaction**, v. 3, CSCW, p. 1–24, 2019.

GUASTINI, R. **Interpretar y argumentar.** Madrid: Centro de Estudios Políticos y Constitucionales, 2014.

HILDEBRANDT, M. The New Imbroglio, Living with Machine Algorithms. In: JANSSENS, L. (ed.). **The Art of Ethics in the Information Society**. Amsterdam: Amsterdam University Press, 2016. p. 55–60.

KELSEN, H. **Teoría pura del derecho.** Traducción de Roberto J. Vernengo. México: Universidad Nacional Autónoma de México, 1982.

KENTON, Z.; EVERITT, T.; WEIDINGER, L.; GABRIEL, I.; MIKULIK, V.; IRVING, G. Alignment of Language Agents. **arXiv preprint**, arXiv:2103.14659, 2021.

KOZLOVSKI, A. Reasons Underdetermination in Meaningful Human Control. **Ethics and Information Technology**, v. 27, n. 59, 2025.

KUMARAGE, P. M.; SAARELA, M. Explainable Generative AI: A Two-Stage Review of Existing Techniques and Future Research Directions. **AI**, v. 7, n. 1, p. 31, 2026.

LIPTON, Z. C. The Mythos of Model Interpretability. **Communications of the ACM**, v. 61, n. 10, p. 36–43, 2018.

LUHMANN, N. *Social Systems*. Stanford: Stanford University Press, 1995.

MAHROUK, A. Epistemic Limits of Local Interpretability in Self-Modulating Cognitive Architectures. **Frontiers in Artificial Intelligence**, v. 8, 2025.

MELLO, M. M.; GUHA, N. Understanding Liability Risk from Using Health Care Artificial Intelligence Tools. **New England Journal of Medicine**, v. 390, n. 3, p. 271–278, 2024.

O'NEIL, C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown, 2016.

PAGALLO, U. *The Laws of Robots: Crimes, Contracts, and Torts*. Dordrecht: Springer, 2013.

PARASURAMAN, R.; RILEY, V. Humans and Automation: Use, Misuse, Disuse, Abuse. **Human Factors**, v. 39, n. 2, p. 230–253, 1997.

PAREDES, F. I. Judicial Discretion, Reasonableness and Judicial Review. **Cuestiones Constitucionales**, n. 42, p. 251–270, 2020.

PRICE, W. N. II. Medical Malpractice and Black-Box Medicine. In: COHEN, I. G.; LYNCH, H. F.; VAYENA, E.; GASSER, U. (eds.). **Big Data, Health Law, and Bioethics**. Cambridge: Cambridge University Press, 2018. p. 295–306.

PRIERGAARD, J. Not My Debt: The Institutional Origins of Robodebt. **Australian Journal of Public Administration**, ago. 2024.

REMUS, D.; LEVY, F. Can Robots Be Lawyers? Computers, Lawyers, and the Practice of Law. **Georgetown Journal of Legal Ethics**, v. 30, p. 501–558, 2017.

SALVI, M.; SEONI, S.; CAMPAGNER, A.; GERTYCH, A.; ACHARYA, U. R.; MOLINARI, F.; CABITZA, F. Explainability and Uncertainty: Two Sides of the Same Coin for Enhancing the Interpretability of Deep Learning Models in Healthcare. **International Journal of Medical Informatics**, v. 197, p. 105846, 2025.

SANTONI DE SIO, F.; MECACCI, G.; CALVERT, S.; HEIKOOP, D.; HAGENZIEKER, M.; VAN AREM, B. Realising Meaningful Human Control Over Automated Driving Systems: A Multidisciplinary Approach. **Minds and Machines**, v. 33, p. 587–611, 2023.

SANTONI DE SIO, F.; VAN DEN HOVEN, J. Meaningful Human Control over Autonomous Systems: A Philosophical Account. **Frontiers in Robotics and AI**, v. 5, p. 15, 2018.

SCHERER, M. U. Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies. **Harvard Journal of Law & Technology**, v. 29, n. 2, p. 353–400, 2016.

SOLUM, L. B. Legal Personhood for Artificial Intelligences. **North Carolina Law Review**, v. 70, n. 4, p. 1231–1287, 1992.

SRIVASTAVA, S. Large Language Models Threaten Language's Epistemic and Communicative Foundations. In: **Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing**. Miami, Florida: Association for Computational Linguistics, 2025. p. 1–15.

SUKSI, M. Administrative Due Process When Using Automated Decision-Making in Public Administration: Some Notes from a Finnish Perspective. **Artificial Intelligence and Law**, v. 29, p. 87–110, 2021.

TEUBNER, G. Digital Personhood? The Status of Autonomous Software Agents in Private Law. **Ancilla Iuris**, p. 107–149, 2018.

TRITTO, P. P.; TORRES, I. C. Jurists of the Gaps: Large Language Models and the Quiet Erosion of Legal Authority. **Masaryk University Journal of Law and Technology**, v. 19, n. 2, p. 179–205, 2025.

UNIÓN EUROPEA. Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo de 13 de junio de 2024 por el que se establecen normas armonizadas en materia de inteligencia artificial. **Diario Oficial de la Unión Europea**, L 2024/1689, 12 jul. 2024.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, Ł.; POLOSUKHIN, I. Attention Is All You Need. In: **Advances in Neural Information Processing Systems 30**, 2017. p. 5998–6008.

VELUWENKAMP, H. Reasons for Meaningful Human Control. **Ethics and Information Technology**, v. 24, n. 51, 2022.

WACHTER, S.; MITTELSTADT, B.; FLORIDI, L. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. **International Data Privacy Law**, v. 7, n. 2, p. 76–99, 2017.

WEI, J.; TAY, Y.; BOMMASANI, R.; RAFFEL, C.; ZOPH, B.; BORGEAUD, S.; YOGATAMA, D. et al. Emergent Abilities of Large Language Models. **Transactions on Machine Learning Research**, 2022.

YEUNG, K. 'Hypernudge': Big Data as a Mode of Regulation by Design. **Information, Communication & Society**, v. 20, n. 1, p. 118–136, 2017.

ZALNIERIUTE, M.; MOSES, L. B.; WILLIAMS, G. The Rule of Law and Automation of Government Decision-Making. **Modern Law Review**, v. 82, n. 3, p. 425–455, 2019.

FERNANDO A. RAMOS-ZAGA

Docente investigador, su trabajo académico se centra en la bioética, la filosofía del derecho, la regulación de nuevas tecnologías y la economía conductual, áreas en las que desarrolla enfoques interdisciplinarios orientados al análisis de las implicaciones sociales, éticas y normativas de la innovación científica y tecnológica. Entre sus publicaciones más recientes destacan los artículos "Hacia una bioética de segunda generación: fundamentos normativos para la edición genética de la línea germinal humana" (En-claves del pensamiento, 2026) y "Reconceptualizing Human Authorship in

the Age of Generative AI: A Normative Framework for Copyright Thresholds" (Laws, 2025).

Dirección profesional: Av. Nicolas Ayllón 8510, Ate, Lima, Perú.

Orcid id: https://orcid.org/0000-0001-6301-9460

Correo electrónico: fernandozaga@gmail.com

Recibido el: 03/10/2025
Aceptado el: 10/05/2026
Publicado el: 09/11/2026

Editor responsable
José Sérgio da Silva Cristóvam
Thanderson Pereira de Sousa

Disponibilidad de los Datos

La autoría declara que todos los datos utilizados en la investigación se encuentran disponibles en un repositorio público, de conformidad con las prácticas de ciencia abierta. La Revista Sequência fomenta el intercambio de datos de investigación que garanticen la transparencia, la reproducibilidad y la verificación de los resultados publicados, respetando, no obstante, los principios éticos aplicables. Por ello, no se exige la divulgación de información que permita la identificación de los sujetos de investigación o comprometa su privacidad. El intercambio de datos debe, por tanto, priorizar la integridad científica y la protección de los datos sensibles, garantizando la difusión de los resultados sin exposición indebida de los participantes.

Este trabajo está bajo una licencia Creative Commons Attribution 4.0 International License.